AFCCC/TN-96/011

AIR WEATHER SERVICE

Consolidated Statistical Background Papers

By
Charles R. Coffin

Air Force Combat Climatology Center

Past Weather Our Future

NOVEMBER 1996

Air Force Combat Climatology Center
859 Buchanan Street
Scott Air Force Base, Illinois 62225-5116

REVIEW AND APPROVAL STATEMENT

AFCCC/TN—96/011, *Consolidated Statistical Background Papers*, November 1996, has been reviewed and is approved for public release. There is no objection to unlimited distribution of this document to the public at large, or by the Defense Technical Information Center (DTIC) to the National Technical Information Service (NTIS).

Larry J White

LARRY J. WHITE, Maj, USAF
Chief, Systems Division

Charles R. Coffin

CHARLES R. COFFIN
Author/Statistician

FOR THE COMMANDER

James S Perkins

JAMES S. PERKINS
Scientific and Technical Information
Program Manager
15 November 1996

REPORT DOCUMENTATION PAGE

2. Report Date: November 1996

3. Report Type: Technical Note

4. Title: Consolidated Statistical Background Papers

6. Authors: Charles R. Coffin

7. Performing Organization Name and Address: Air Force Combat Climatology Center (AFCCC/SYT), 859 Buchanan St., Scott AFB IL 62225-5116

8. Performing Organization Report Number: AFCCC/TN—96/011

12. Distribution/Availability Statement: Approved for public release; distribution is unlimited.

13. Abstract: This technical note is a compilation of several years' worth of background papers covering a wide range of topics in statistics. Many sample SAS procedures are also included.

14. Subject Terms: CONSOLIDATED STATISTICAL BACKGROUND PAPERS, PERIOD OF RECORD, FREQUENCY DISTRIBUTION, SKEWNESS, KURTOSIS, CORRELATION, REGRESSION CANONICAL CORRELATION, COEFFICIENTS, CONTINGENCY TABLE, CHI-SQUARE TEST, CRAMER'S V, JOINT PROBABILITY, MARGINAL PROBABILITY, CONDITIONAL PROBABILITY, ODDS RATIO, WEIGHTED LEAST SQUARES, MEAN, MEDIAN, MODE, INFERENTIAL STATISTICS, STANDARD DEVIATION, COEFFICIENT OF VARIATION, T-TEST, L-MOMENT, DISCORDANCY MEASURE, PERCENTILES, TETRACHORIC CORRELATION, TRANSNORMALIZATION, AUTOCORRELATION, COEFFICIENT OF DETERMINATION, MULTICOLLINEARITY, REGRESSION ANALYSIS, RIDGE REGRESSION, DISCRIMINANT ANALYSIS, HEIDKE SKILL SCORE, CENTRAL LIMIT THEOREM, MEAN SQUARE ERROR, BRIER SCORE

15. Number of Pages: 119

17. Security Classification of Report: Unclassified

18. Security Classification of this Page: Unclassified

19. Security Classification of Abstract: Unclassified

20. Limitation of Abstract: UL

Standard Form 298

PREFACE

The Air Force Combat Climatology Center's statistician, Mr Charles Coffin, regularly writes background papers covering a wide range of topics in statistics. Many of these papers also include samples of statistical analysis procedures (SAS) often used at AFCCC. The goal of these papers is to keep AFCCC analysts up to date on current statistical techniques.

Mr Kevin Havener (as Capt Havener of AFCCC/SYT) and Mr Anthony Warren (as Capt Warren of AFCCC/SYT) made comments on drafts of different chapters of this technical note. Mr Warren was helpful with Word Perfect. Mr Havener was helpful in converting text and equations from Word Perfect to Miscrosoft Word. The AFCCC publications services team (Mr Gene Newman, Ms Kristine Byrnside, and SSgt Le La Hartman) edited the technical note.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

PERIOD OF RECORD

1.1 Introduction. Do you have enough data to accurately summarize the true climate of an area? According to the World Meteorological Organization (WMO) publication, WMO-NO.208.TP.108 (1967), determining the period of record (POR) necessary to calculate mean climatological values has perplexed climatologists for a long time. Air Force Combat Climatology Center receives frequent taskings to provide worldwide climatological summaries and statistics for both planning and contingency support. Often data requests are for locations with erratic reporting practices. There is no simple answer to this question, but this chapter attempts to provide analysts guidance on how to decide if there is sufficient data to meet their needs.

1.2 Discussion.

a. The WMO states a longer POR is needed for elements at sites with considerable climatic fluctuations than for those with little climatic fluctuations. Elements such as temperature, humidity, and cloud amount, which show less variability, do not require as many years of observations as precipitation, which is much more variable. Mountainous stations require more years of data than stations in the plains.

b. The WMO generally accepts that a POR of 10 years provides sufficient data for most common climatological elements. Landsberg and Jacobs (1951), state that the proper period of record varies from element to element, from season to season, and from region to region. Table 1-1 is extracted from that publication.

c. Brooks and Carruthers (1953) point out that the whole of climatology is based on the study of samples. A sample is a group of one or more observations. According to Brooks and Carruthers, a sample of 30 or more independent observations may be regarded as a large sample. However, individual observations of most weather variables are serially correlated. To get a random sample of independent observation values, you can spread observations several days apart. You can also reduce serial correlation by using a sample consisting of monthly means instead of individual observations. If you have a random sample of 100 or more means, you can get a good grasp of the underlying distribution of variable means.

If there are not enough observations to calculate a certain mean, then it is best to combine groupings, e.g., calculate seasonal rather than monthly means. It is best to study a POR at first in general detail,

Table 1-1. Recommended period of record (in years) for various regions (ET=Extra Tropical, T=Tropical).

Climatic element	Islands		Shore		Plains		Mountains	
	ET	T	ET	T	ET	T	ET	T
Temperature	10	5	15	8	15	10	25	15
Humidity	3	1	6	2	5	3	10	6
Cloud	4	2	4	3	8	4	12	6
Visibility	5	3	5	3	5	4	8	6
Precipitation	25	30	30	40	40	40	50	50

then combine data into different groups that prove to be most convenient. There is no real loss of accuracy in combining when there are so few observations in a particular group. The number of bins (or classes in SAS), into which a series of observations should be divided, depends on the situation. Brooks and Carruthers give the following rough guide: the number of bins should be no more than five times the logarithm of the total number of observations. If you have 100 (or 10^2) observations, you should not have more than (2 X 5) = 10 bins.

The best way to make a preliminary study of an observation series, according to Brooks and Carruthers, is to form a frequency distribution. Percentiles on all the monthly means gives information about the median (50th percentile) of all the means and the extreme mean percentiles (1st and 99th). The 1st percentile of means indicates the lowest extreme mean value and the 99th the highest.

d. Climatologists currently define the term "normal" to be a mean of a climatic element over a time period of 30 years, comprising at least three consecutive 10-year periods. Panofsky and Brier (1958) claim that a mean based on 15 years of data gives the best estimate for next year's mean and is therefore preferable to climatic normals based on more than 15 years. Court (1968) recommends using climatic normals over a 15-year period of record, rather than a 30-year period of record, with recomputation of the normals every 5 years. He states the median values based on 15 years is an even better predictor than the mean value. Finally, Court points out that 7 years is a suitable time period for defining the climate of a region.

e. You should consider means with equal samples in the construction of climatological normals, according to Panofsky and Brier (1958). For example, suppose you want to know the grand mean January 12Z temperature. You have 20 years of mean January 12Z data, however, 15 years of the January means were calculated using 31 daily values while the means for the remaining 5 years are based on 10 or less daily values. In this case, you really only have a 15 year POR, not 20. Statisticians usually design a comparison study to ensure the number of observations is the same in each data sample. Equal sample sizes are preferred since they are simpler to analyze as well as more efficient. Comparison of means with unequal sample sizes is more sensitive to the violation of statistical assumptions. You gain more by comparing means of equal sizes than by arguing that a statistic has to be based on a set number of data points.

f. If some data is missing, Panofsky and Brier suggest that a mean computed from a short record can be augmented by the use of information at surrounding sites. WMO-NO.208.TP.108 (1967) states that missing data can be dealt with in three ways:

(1) Check the available climatological data at different stations and use a site with a better relative data consistency;

(2) Interpolate missing data in a recorded period;

(3) Reconstruct insufficiently long climatological series by referring to data of a neighboring comparable station for which sufficiently long data series is available.

1.3 Conclusion. Although the WMO recommends using a POR of between 10 and 15 years for a valid climatological study, some climatologists lean toward using between 7 to 15 years. Percentiles on monthly mean values show us the breakdown of mean values. A cumulative frequency distribution on a data sample gives us a more complete picture of the data sample. If you have enough data, then it might be a good idea to separate the frequency distribution into two 10-year periods of record. If the two periods of record generate the same frequency distribution, then you know a 10-year POR is sufficient. In addition, according to Panofsky and Brier (1958), you should compare means with equal sample sizes. If some data are missing, then WMO (WMO-NO.208.TP.108) suggests several methods to fill in the data without inducing very much uncertainty.

Chapter 2

PERIOD OF RECORD USING CUMULATIVE FREQUENCY DISTRIBUTION

2.1 Introduction. The focus of this chapter is on a strategy that may be used to determine the minimum length of time needed to obtain a "satisfactory" period of record. This strategy is taken from Air Weather Service Technical Report 105-25. SAS code is also provided showing how this technique can be employed.

2.2 Discussion.

a. The purpose of AWS/TR 105—25 is to establish the number of years that are needed in order to obtain a relatively constant frequency distribution for a given meteorological element. A constant frequency distribution is attained when the addition of years of record to the database does not significantly alter the interpretation of climatic records. Once the number of years necessary to achieve this result is established, the work of compiling climatic records can be greatly streamlined.

b. The AWS study describes the use of the "cumulative frequency distribution" to establish the minimum satisfactory length for a POR. The cumulative frequency distribution shows the number of observations of a given element for each class (such as "year"), up to and including the most current class. Table 2-1 shows the cumulative frequency for January cloud cover, divided by tenths into four intervals. Table 2-2 (see next page) shows the cumulative frequencies for October cloud cover data, over a longer period.

As seen in Table 2-1, after 3 years none of the cumulative percentage frequencies in any step changes by more than five percent with the addition of subsequent years. Table 2-2 shows that after 6 years none of the cumulative percentage frequencies in any step changes by more than five percent with the addition of subsequent years. This holds true even if data collection is extended to 35 years (through 1906). These results indicate 6 years' worth of data constitute a satisfactory POR for the study of October cloud cover frequencies in the eastern United States, while 3 years may be sufficient for January cloud cover frequencies. Note: use of 5 percent as the constancy limit is arbitrary. The data must be divided into more than one interval for the AWS technique to be applied.

Table 2-1. Frequencies of occurrence of cloud cover intervals (in tenths) for Washington D.C., January.

Year	Frequencies in days (individual years)				Cumulative frequencies (years added)				Cumulative percentage frequencies			
	0-2	3-5	6-8	9-10	0-2	3-5	6-8	9-10	0-2	3-5	6-8	9-10
1940	13	6	2	10	13	6	2	10	41.9	19.4	6.5	40.3
1939	10	2	4	15	23	8	6	25	37.1	12.9	9.7	40.3
1938	13	1	2	15	36	9	8	15	38.7	9.7	8.6	43.0
1937	6	1	5	19	42	10	13	59	33.9	8.0	10.5	47.6
1936	13	2	3	13	55	12	16	72	35.5	7.7	10.3	46.4
1935	17	1	2	11	72	13	18	83	38.7	7.0	9.7	42.9

Table 2-2. Cumulative percentage frequencies of cloud cover intervals (in tenths) for Washington D.C., October.

	0-2	3-5	6-8	9-10
1940	48.39	9.68	3.22	38.71
1939	46.77	6.45	4.84	41.90
1938	56.99	5.38	4.30	33.33
1937	58.06	4.84	4.84	32.26
1936	56.77	5.81	7.10	30.32
1935	56.45	6.45	8.60	28.49
1934	58.53	6.91	7.83	26.73
1933	60.89	6.05	7.26	25.80
1932	59.86	5.38	6.81	27.95
1931	60.00	4.84	7.10	28.06
.....	.....	.....	.....	.....
1906	58.34	6.82	9.12	25.71

2.3 Example. The SAS statements shown below can be used to generate the results shown in Table 2-1.

```
*READ IN CLOUD DATA FOR YEAR;
DATA ONE;
INPUT YEAR CLOUD;
CARDS;
1991 0
1991 4
1991 7
1991 10
1992 3
.
.
.
1993 7
;
RUN;


*DIVIDE CLOUD DATA INTO INTERVALS;

DATA TWO CLD02 CLD35 CLD68 CLD910;
SET ONE;
IF CLOUD GE 0 AND CLOUD LE 2 THEN
OUTPUT CLD02;
IF CLOUD GE 3 AND CLOUD LE 5 THEN
OUTPUT CLD35;
IF CLOUD GE 6 AND CLOUD LE 8 THEN
OUTPUT CLD68;
IF CLOUD GE 9 AND CLOUD LE 10 THEN
OUTPUT CLD910;
RUN;

*CALCULATE FREQUENCIES;

DATA THREE;
SET TWO;
PROC FORMAT;
VALUE CLD 0-2=’0 TO 2'
          3-5=’3 TO 5'
          6-8=’6 TO 8'
          9-10=’9 TO 10';
PROC MEANS DATA=CLD02 N NOPRINT;
FORMAT CLOUD CLD.;
OUTPUT OUT=FCLD02 N=FCLD02;
CLASS YEAR;
PROC MEANS DATA=CLD35 N NOPRINT;
FORMAT CLOUD CLD.;
OUTPUT OUT=FCLD35 N=FCLD35;
CLASS YEAR;
PROC MEANS DATA=CLD68 N NOPRINT;
FORMAT CLOUD CLD.;
OUTPUT OUT=FCLD68 N=FCLD68;
CLASS YEAR;
PROC MEANS DATA=CLD910 N NOPRINT;
FORMAT CLOUD CLD.;
OUTPUT OUT=FCLD910 N=FCLD910;
CLASS YEAR;
RUN;
DATA FOUR;
MERGE FCLD02 FCLD35 FCLD68 FCLD910;
IF YEAR = . THEN DELETE;
KEEP FCLD02 FCLD35 FCLD68 FCLD910
YEAR;
RUN;

*CALCULATE CUMULATIVE FREQUENCIES;

DATA FIVE;
SET FOUR;
CFCLD02 + FCLD02;
CFCLD35 + FCLD35;
CFCLD68 + FCLD68;
CFCLD910 + FCLD910;
RUN;

*CALCULATE CUMULATIVE PERCENTAGE
FREQUENCIES;

DATA SIX;
SET FIVE;
CPCLD02 = CFCLD02/(CFCLD02 + CFCLD35 +
CFCLD68 + CFCLD910);
CPCLD35 = CFCLD35/(CFCLD02 + CFCLD35 +
CFCLD68 + CFCLD910);
CPCLD68 = CFCLD68/(CFCLD02 + CFCLD35 +
CFCLD68 + CFCLD910);
CPCLD910 = CFCLD910/(CFCLD02 + CFCLD35
+ CFCLD68 + CFCLD910);
CPCLD02 = CPCLD02*100;
CPCLD35 = CPCLD35*100;
CPCLD68 = CPCLD68*100;
CPCLD910 = CPCLD910*100;
RUN;

PROC PRINT;
RUN;
```

Chapter 3

A NOTE ON CLIMATOLOGICAL NORMALS

3.1 Introduction. This chapter discusses the problem of determining the length of the reference period necessary for the calculation of climatological mean values. Solutions suggested by several authors are presented.

3.2 Discussion.

a. Use of the term "normal" to describe the mean of a long series of observations first appeared in an 1840 article by Dove. The general public believes the "normal" is the most frequent value. Climatologists understand that a normal value is the long term mean. This is usually not the most frequent value (mode), nor the value above which half the cases fall (median).

b. Normals have been used for two purposes—comparison and prediction. The normal serves as a reference value by which to compare past and present values. For example, at St. Charles, Mo., the normal or mean temperature (30-year period, 1951-1980) for January is 28.8° F. If we have an actual mean January 1990 temperature of 19.0 degrees, then we know this is 9.8 degrees colder than normal. As predictors, climatic normals are often inefficient estimators of future conditions. Normals provide only one measure, they tell little about climatic change, nonrandom fluctuations (trends) or extremes.

c. In 1956, the World Meteorological Organization (WMO) recommended that data from the most recent 30 year period at a given location be used in the calculation of climatological normals. This decision has frequently been critically examined. Due to climatic fluctuations, the statistics based on a 30-year, or even a 50-year period of record may not be so absolutely stable throughout the world as to be termed "normal" for all locations. Climatic fluctuations vary in magnitude in various parts of the world. A POR, which is sufficient to provide a representative measure of conditions at one location, may be insufficient at another. Compounding this problem is the fact that as the POR expands, maintaining homogeneity of the data becomes more difficult. Climatological statistics obtained from too long a period may not be representative of contemporary conditions.

d. Rubinstein (1962), Kuznetsova (1964), and Shvec (1964) have addressed this problem, studying the elements temperature, humidity, wind, and radiation. They found that for temperature data, 10-year monthly averages may vary by as much as 10° C between two decades. For precipitation data, 30-year periods of record are inadequate for the purpose of obtaining stable average monthly precipitation values. For humidity records, data for 30 to 35 years provide sufficiently stable average values. Estimates of the mean wind speed, as well as maximum wind speed occurring once in 10, 20, or 50 years, can be calculated with a sufficient degree of accuracy on the basis of 20 to 25 years' worth of data. Radiation characteristics can be calculated on the basis of 25 to 30 years.

e. The U.S. Army Air Force (1943) also conducted a study of the length of record needed to obtain satisfactory climatological summaries. An attempt was made to find the number of years needed to yield a relatively constant frequency distribution (within 5 percent) for visibility, cloudiness, cloud height, wind speed, and precipitation. It was found that data for about 7 to 10 years was needed for visibility, cloud, and wind, while for precipitation the required POR was about 20 years.

f. Regarding the number of years needed to obtain stable frequency distributions, Landsberg and Jacobs (1951) have also indicated that the number of years varies from element to element, from season to season, and from region to region. Their findings are summarized in Table 3-1 on the next page.

Table 3-1. Approximate number of years needed to obtain stable frequency distribution. (ET = Extratropical, T = Tropical).

Climatic element	Islands		Shore		Plains		Mountains	
	ET	T	ET	T	ET	T	ET	T
Temperature	10	5	15	8	15	10	25	15
Humidity	3	1	6	2	5	3	10	6
Cloud	4	2	4	3	8	4	12	6
Visibility	5	3	5	3	5	4	8	6
Precipitation	25	30	30	40	40	40	50	50

3.3 Conclusion. Studies have shown that a shorter POR is needed to obtain stable estimates of "normal" values for less variable elements such as temperature, humidity, and cloud amount. More variable element, such as precipitation, require longer periods of record. Also, more data is required for stations in mountainous areas than for stations in the plains. It's very difficult to specify a uniform period that can be used as a reference period for all elements, and which can be considered representative for the world as a whole.

Chapter 4

SIGNIFICANT FIGURES

4.1 Introduction. There is widespread belief that the accuracy of a measurement or computed result is indicated by the number of decimal places required to express it. This is erroneous. Instead, the accuracy is indicated by the number of significant figures in the result. This section presents definitions and examples of significant figures, truncation errors, and round-off errors.

4.2 Discussion.

a. Truncation error is the difference between the "true" answer and the answer obtained by a mathematical calculation. Often, a mathematical expression is solved numerically using finite approximations of infinite expressions (e.g., the evaluation of integrals). The difference between the true answer and the answer obtained from the finite process is known as the truncation error. Truncation error is under the programmer's control.

b. Round-off error is the error due to dropping off digits (e.g., approximating 1/3 with 0.333 results in a round-off error of 0.0003333333...). Round-off errors accumulate with increasing calculations. Round-off errors are a characteristic of computer hardware, but may be lessened by choosing algorithms that do not magnify it unnecessarily.

c. Round-off error can be better understood by relating it to the number of significant figures, which is the number of digits in the answer whose values we are reasonably sure of. To round off a number to fewer significant digits than were specified originally, truncate the number as desired and treat the excess digits as a decimal fraction (according to Bevington and Robinson, 1992). If the fraction is greater than one-half, increment the new least significant digit. If the fraction is less than one-half, do not increment. If the fraction equals one-half, increment the least significant digit only if it is odd.

d. Rules for determining the number of significant figures:

1) All nonzero digits are significant (e.g., 159.75 contains 5 significant digits).

2) All zeros between two nonzero digits are significant (e.g., 108.005 contains 6 significant figures).

3) Unless otherwise indicated, all zeroes to the left of an understood decimal point, but to the right of a nonzero digit are not significant. The concept of an understood decimal point is best illustrated by using scientific notation. For example, if we express the number 202,000 by 2.02 x 10^5, the measurement has three significant digits.

4) All zeroes to the left of expressed decimal points and to the right of a nonzero digit are significant. Expressing the previous example as 202,000 (2.02000 x 10^5) results in six significant digits.

5) All zeros to the right of a decimal point, but to the left of a nonzero digit are not significant (e.g., 0.000647 contains three significant figures).

6) All zeros to the right of a decimal point and to the right of a nonsignificant digit are significant (e.g., 0.07080 and 20.00 each contains four significant figures).

e. To determine the number of significant figures when adding and subtracting, first round all measurements to the accuracy of the least accurate measurement, then add or subtract.

1) Addition example. Add 17.35, 25.6, and 8.498. The value with the least accuracy is 25.6, known only in tenths. Rounding the other measurements to tenths yields (17.4 + 25.6 + 8.5), which equals 51.5.

2) Subtraction example. Subtract 36.8 from 97. Round 36.8 to 37. Subtracting 37 from 97 yields 60. The result is expressed as 60. (not 60) to indicate that the zero is significant (i.e., the final result has two significant figures).

f. When multiplying or dividing a group of numbers, use the following rule: If no number in the group contains fewer than *s* significant figures, the others should be rounded off, if necessary, to $s + 1$ significant figures. After all calculations are carried out, the result is rounded off to *s* significant figures. Consider the following example to illustrate this point.

4.2.1 Example. Suppose you want to determine the volume of a cylinder ($V = [4/3]pr^2h$). The radius (*r*) of the cylinder is measured to as 22.264 cm, and the height (*h*) is 7.2 cm. In this example the value of *s* is 2, so the radius should be rounded to three significant figures ($s + 1 = 3$). This gives a radius of 22.3. Plugging these values into the expression for the volume produces a result of 14,990 cm^3. Rounding this result gives 1.50 x 10^4 cm^3. Since the final result can have no more than *s* significant figures, the volume should be reported as 1.5 x 10^4 cm^3.

4.3 Conclusion. It's pointed out in the literature that the rules of significant figures should be applied with common sense. The rules are only guidelines and there are exceptions. SAS carries its calculations out to an excessive number of decimal places. According to Sachs (1984), mean values and standard deviations should not be stated with more than two decimal places more than the original data. This is appropriate when the sample size is large. Dimensionless constants like skewness, kurtosis, correlation, and regression coefficients should be stated with at most four significant figures.

Chapter 5

CONTINGENCY MEASURES AND MEASURES OF ASSOCIATION

5.1. Introduction. Analysts are often interested in testing the similarity of two frequency distributions. A typical question might be, "Is the cloud-cover frequency distribution at point A similar to that at point B?" Contingency tables provide a useful technique for studying the relationships among two or more variables. This chapter describes the construction of contingency tables, and discusses some of the measures of association that can be derived from the information in contingency tables.

5.2. Discussion.

a. A contingency table (also referred to as a cross-tabulation) depicts the joint distribution of two or more variables. Tables 5-1 and 5-2 show examples of contingency tables.

b. In the Table 5-2, the wind direction, and the month are variables. The classes within the variable "month" are June, July, and August. The classes within the variable "wind speed" are $\leq$ 25 Km/hr, and > 25 Km/hr. The first class of the wind direction is S and SW. The second class is W, NW, N, and Calm. The third class is all directions. The numbers such as 13, 8,8, under S and SW represent the frequencies, or numbers of occurrences within the total sample (n = 371). The numbers in parentheses represents the "expected" number of occurrences, which are calculated as shown.

$$\text{Expected Value} = \frac{(\text{row total})(\text{column total})}{\text{total sample size}}.$$

For example, the expected value of 11, for S and SW winds, is calculated as follows:

$$\text{Expected Value (11)} = \frac{(29)(145)}{371}.$$

Table 5-1. General contingency table format.

	Y_1	Y_2	...	Y_n	Row Total
X_1	O_{11}	O_{12}	...	O_{1n}	A_1
X_2	O_{21}	O_{22}	...	O_{2n}	A_2
...	...	...	...	...	...
X_m	O_{m1}	O_{m2}	...	O_{mn}	A_m
Column Total	B_1	B_2	...	B_n	N

Table 5-2. Sample contingency table, showing a comparison of the frequency distributions of surface winds, in 3 separate months.

	$\leq$ 25 Km/hr		> 25 Km/hr	
	S and SWN,	W, NW, N, and Calm	All directions	Total
June	13 (11)	91 (101)	41 (35)	145
July	8 (9)	74 (76)	28 (25)	110
August	8 (9)	93 (81)	15 (26)	116
Total	29	258	84	371

c. The data in contingency tables can be used to calculate various types of measures of association that describe the relationships between variables. Some of these measures of association are describe below.

(1) The Chi-Square test. Using the observed and expected values from the contingency table shown above, the chi-square test can be used to determine if two variables are independent of each other, such as the month and the wind speed. The chi-square statistic is

$$\chi^2 = \Sigma \frac{(observed\ value\ -\ expected\ value)^2}{expected\ value}.$$

"Small" values of the chi-square statistic indicate the absence of a relationship between the variables, meaning the variables are statistically independent. A "large" chi-square statistic implies a relationship exists between the variables. Chi-square tables are used to determine whether the values are "small" or "large," at various probability levels for various degrees of freedom. In our example the chi-square value is 10.35, which is calculated as follows:

$$\chi^2 = \frac{(13 - 11)^2}{11} + \frac{(8 - 9)^2}{9} + \frac{(8 - 9)^2}{9} + \frac{(91 - 101)^2}{101}$$

$$+ \frac{(74 - 76)^2}{76} + \frac{(93-81)^2}{81} + \frac{(41 - 33)^2}{33} + \frac{(28 - 25)^2}{25} + \frac{(15-26)^2}{26}$$

$$= 10.35.$$

To use a chi-square table, one must first determine the degrees of freedom, v. In the present example

$$v = (rows - 1)(columns - 1) = 2 \times 2 = 4.$$

In this case, c^2 = 10.35. The correct interpretation of the test is that if you sample repeatedly from a chi-square distribution with 4 degrees of freedom, you will calculate an $c^2 \geq 10.35$ only three times out of 100 on the average. Obviously, this chi-square value is "large." Thus, the month and wind speed are related.Chi-square is sensitive to sample size, and is therefore not appropriate as a measure of strength of relationship. To measure strength of relationship, the dependence of chi-square on sample size must be eliminated.

(2) The SAS FREQ procedure will generate the chi-square test, as well as other measures of the strength of relationship such as the contingency coefficient and the phi coefficient, which are calculated as shown below.

$$Contingency\ Coefficient\ = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

$$Phi\ Coefficient\ (\Phi) = \sqrt{\frac{\chi^2}{n}}.$$

The contingency coefficient and the phi coefficient both have a value of zero when chi-square is zero (when no association exists between the two variables). The contingency coefficient has a lower bound of zero, attained when chi-square is zero. The contingency coefficient is always less than one, but the upper bound approaches one as the size of the contigency table increases (Shulman, 1992). The limits of the phi coefficient are -1 ≤ phi ≤ +1 for a 2 x 2 contigency table and 0 ≤ phi ≤ +1 otherwise. The phi coefficient is usually used only for tables with two rows and/or two columns since its upper bound may be larger than one for larger tables.

(3) Cramer's V. Cramer's V is used to measure the strength of relationships for tables with more than two rows and two columns. It is calculated as shown below

$$V = \frac{\Phi}{\sqrt{\min(r\ -\ 1, c\ -\ 1)}}.$$

In this formula, min (r-1, c-1) stands for the smaller of (# of rows - 1) and (# of columns -1); F is the phi coefficient. Cramer's V has a range from zero to one, with a value of zero indicating no association exists between variables.

Chapter 6

PROBABILITY

6.1 Introduction. Probability is usually thought of as the fraction of time that an event **A** will occur, in the total set of observations. The terms joint, marginal, and conditional probability are common terms. This chapter provides definitions of the various types of probability, and gives an example of how probability can be used in meteorology.

6.2 Discussion.

a. Probability values can range from 0 to 1. If the probability of a given event is 1 then that event is known as a certain event. If the probability of a given event is 0, it is referred to as an impossible event.

b. Computing Probability. The terms probability and relative frequency are often interchanged. The relative frequency of an event **A** is simply the ratio of the number of occurrences of event **A** (N_A) divided by the total number of all possible events (N). A relative frequency is only an approximation to the probability.

$$P(A) \approx \frac{N_A}{N}\,. \tag{1}$$

Theoretically, this value becomes the exact probability only in the limit as N approaches infinity:

$$P(A) = \lim_{N \to \infty} \frac{N_A}{N}\,. \tag{2}$$

In practice, we use equation (1) to estimate probabilities. However, this introduces an uncertainty that increases as N decreases.

c. Independent and Dependent Events. Two events (**A, B**) are said to be dependent if the probability of occurrence of one event is affected by the occurrence of the other event. Two events (**A, B**) are said to be independent if the probability of occurrence of one event is not affected by the occurrence of the other event.

d. Joint Probability. The set containing the number of cases which belong to both events **A** and **B** is called the intersection of **A** and **B**. The intersection of **A** and **B** is denoted by $(A \cap B)$. The probability of $(A \cap B)$, is given by:

$$P(A \cap B) = \frac{N_{(A \cap B)}}{N} \tag{3}$$

where $N_{(A \cap B)}$ is the total number of events in $(A \cap B)$ and N is the total number of all possible events. The probability of $(A \cap B)$ is referred to as the joint probability of events **A** and **B**, and can be written P(**AB**).

e. Marginal Probability. The marginal probability is the sum of the joint probabilities of a given event. The marginal probability of event $\mathbf{A}_i$ is given by:

$$P(A_i) = \sum_{j=1}^{T} P(A_i B_j) \tag{4}$$

where $P(\mathbf{A}_i\ \mathbf{B}_j)$ represents the joint probability of events $\mathbf{A}_i$ and $\mathbf{B}_j$, with T being the total number of categories.

f. Union. The number of cases which belong to either event $\mathbf{A}_i$ or event $\mathbf{B}_j$, or both $\mathbf{A}_i$ and $\mathbf{B}_j$ is called the union of $\mathbf{A}_i$ and $\mathbf{B}_j$ and is designated $\mathbf{A}_i \cup \mathbf{B}_j$. The probability of $\mathbf{A}_i \cup \mathbf{B}_j$ is given by:

$$P(A_i \cup B_j) = P(A_i) + P(B_j) - P(A_i \cap B_j)\,. \tag{5}$$

g. Conditional Probability. The probability that an event $\mathbf{A}_i$ occurs given that $\mathbf{B}_j$ has occurred is called the conditional probability of $\mathbf{A}_i$ given $\mathbf{B}_j$ and is denoted by $P(\mathbf{A}_i|\mathbf{B}_j)$. The conditional probability is:

$$P(A_i|B_j) = \frac{P(A_i \cap B_j)}{P(B_j)}\,. \tag{6}$$

If $P(\mathbf{B}_j) = 0$ (i.e., $\mathbf{B}_j$ is an impossible event), then the conditional probability of $\mathbf{A}_i$ given $\mathbf{B}_j$ is undefined.

6.3 Example. The following examples may help illustrate the differences between these various terms. When computing probabilities, it is best to start with a contingency table, as shown below.

Event	$\mathbf{B_1}$	$\mathbf{B_2}$	...	$\mathbf{B_n}$
$\mathbf{A_1}$	$n(A_1B_1)$	$n(A_1B_2)$	...	$n(A_1B_n)$
$\mathbf{A_2}$	$n(A_2B_1)$	$n(A_2B_2)$	...	$n(A_2B_n)$
...	...	...	...	...
$\mathbf{A_m}$	$n(A_mB_1)$	$n(A_mB_2)$		$n(A_mB_n)$

a. Consider the following example. Event $\mathbf{A_1}$ is the observation of a halo around the moon. Event $\mathbf{A_2}$ is the complement of this event - no halo is observed. Event $\mathbf{B_1}$ is the occurrence of precipitation with 48 hours. Event $\mathbf{B_2}$ is no precipitation with 48 hours.

	PRECIP ($\mathbf{B_1}$)	**NO PRECIP ($\mathbf{B_2}$)**	**MARGINAL TOTAL**
HALO ($\mathbf{A_1}$)	497	149	646
NO HALO ($\mathbf{A_2}$)	819	819	1638
MARGINAL TOTAL	1316	968	2284

Using the data in this contingency, the following probabilities can be calculated:

b. Joint Probabilities:

$$P(\mathbf{A_1B_1}) = P(\mathbf{A_1} \text{ Ç } \mathbf{B_1}) = 497/2284 = 0.218$$

$$P(\mathbf{A_1B_2}) = 149/2284 = 0.065$$

$$P(\mathbf{A_2B_1}) = 819/2284 = 0.359$$

$$P(\mathbf{A_2B_2}) = 819/2284 = 0.359.$$

c. Marginal Probabilities:

$$P(\mathbf{A_1}) = P(\mathbf{A_1B_1}) + P(\mathbf{A_1B_2}) = 646/2284 = 0.283$$

$$P(\mathbf{A_2}) = 1638/2284 = 0.717$$

$$P(\mathbf{B_1}) = 1316/2284 = 0.576$$

$$P(\mathbf{B_2}) = 968/2284 = 0.424.$$

d. Unions. The probability $P(\mathbf{A_1} \text{ È } \mathbf{B_1})$ is the probability of having either a halo around the moon or rain within 48 hours:

$$P(\mathbf{A_1} \text{ È } \mathbf{B_1}) = P(\mathbf{A_1}) + P(\mathbf{B_1}) - P(\mathbf{A_1} \text{ Ç } \mathbf{B_1}) = 0.283 + 0.576 - 0.218 \qquad = 0.642$$

$$P(\mathbf{A_2} \text{ È } \mathbf{B_1}) = 0.717 + 0.576 - 0.359 = 0.935$$

$$P(\mathbf{A_1} \text{ È } \mathbf{B_2}) = 0.283 + 0.424 - 0.065 = 0.641$$

$$P(\mathbf{A_2} \text{ È } \mathbf{B_2}) = 0.717 + 0.424 - 0.359 = 0.782.$$

e. Conditional Probabilities:

$$P(A_1|B_1) = P(A_1B_1)/P(B_1) = 0.218/0.577 = 0.378$$

$$P(A_2|B_1) = P(A_2B_1)/P(B_1) = 0.359/0.577 = 0.622$$

$$P(A_1|B_2) = P(A_1B_2)/P(B_2) = 0.065/0.423 = 0.154$$

$$P(A_2|B_2) = P(A_2B_2)/P(B_2) = 0.359/0.423 = 0.849$$

$$P(B_1|A_1) = P(A_1B_1)/P(A_1) = 0.218/0.283 = 0.770$$

$$P(B_2|A_1) = P(A_1B_2)/P(A_1) = 0.065/0.283 = 0.230$$

$$P(B_1|A_2) = P(A_2B_1)/P(A_2) = 0.359/0.717 = 0.501$$

$$P(B_2|A_2) = P(A_2B_2)/P(A_2) = 0.359/0.717 = 0.501.$$

f. Interpretation. The probability of observing a halo around the moon, $P(A_1)$, is 0.283 (28.3 percent). The probability of rain occurring within any given 48-hour period, $P(\mathbf{B}_1)$, is 0.577. The probability of having both a halo around the moon and rain within a 48-hour period, $P(\mathbf{A}_1\mathbf{B}_1)$, is 0.218. The probability of having rain occur within 48 hours given you see a halo around the moon, $P(\mathbf{B}_1|\mathbf{A}_1)$, is 0.770. The probability that a halo was present at a particular time given that rainfall occurred within a subsequent 48-hour period, $P(A_1|\mathbf{B}_1)$, is 0.378. Note the difference between these last three probabilities. In the first one, $P(\mathbf{A}_1\mathbf{B}_1)$, there are no assumptions. This is the unconditional probability of observing both these events. The last two probabilities are conditional. With the second probability, $P(\mathbf{B}_1|\mathbf{A}_1)$, it is given that event $\mathbf{A}_1$ (halo) has already occurred, and this value then represents the probability that event $\mathbf{B}_1$ (rain) will occur. With the last probability, $P(\mathbf{A}_1|\mathbf{B}_1)$, you are given that event $\mathbf{B}_1$ (rain) has occurred and this value then represents that probability that event $\mathbf{A}_1$ (halo) will occur. Thus, one can conclude from this set of data that it is likely rain will occur within 48 hours of seeing a halo (probability of 77 percent). However, most instances of rainfall are not preceded by a halo (only a 38 percent probability that a rain event is preceded by a halo).

Chapter 7

ODDS RATIO

7.1 Introduction. There is often a confusion of probability with odds. For example, a probability of 50 percent has been mistakenly interpreted to mean that the odds were 50 for one event against 100 for the other. If odds were added to the statement of probability, then this difficulty could be avoided. Another measure of association between two variables is the odds ratio. The odds ratio can be used as a decision making tool and can enable one to know the odds of an outcome. This chapter discusses definitions of probability, odds, odds ratio, and how both the SAS FREQ and SAS CATMOD procedures can be used to generate the odds ratio.

7.2 Discussion.

a. Definitions.

(1) **Probability** - Probability (P_1) is based on the ratio of number of favorable events to number of possible events where Q_1 is $1 - P_1$. Q_1 is ratio of number of unfavorable events to number of possible events.

(2) **Odds** - If P_1 is the probability or rate at which an event occurs in the population where $Q_1 = 1 - P_1$, then the odds associated with that event are P_1/Q_1.

(3) **Odds Ratio** - If P_1 is the probability or rate at which an event occurs in the first population where $Q_1 = 1 - P_1$, then the odds associated with that event in the first population are P_1/Q_1 denoted by O_1. Similarly, the odds associated with the event in the second population are P_2/O_2 denoted by O_2. The odds ratio is simply the ratio of these two odds, O_2/O_1 or P_2Q_1/P_1Q_2.

b. Example. In Table 7-1, individual p values are obtained by dividing each of the individual n values by $n_{..}$. The odds ratio is then calculated.

c. Interpretation of odds ratio. Odds ratio of 1.9 indicates that halo events are 1.9 times as likely to be associated with precipitation within 48 hours as no halo events.

Table 7-1. Precipitation within 48 hours.

	Yes	No	Total
Halo	151 (n_{11})	140 (n_{12})	291 ($n_{1.}$)
No Halo	109 (n_{21})	192 (n_{22})	301 ($n_{2.}$)
Total	260 ($n_{.1}$)	332 ($n_{.2}$)	592 ($n_{..}$)
Halo	0.255 (= p_{11})	0.236 (= p_{12})	0.492 (= $p_{1.}$)
No Halo	0.184 (= p_{21})	0.324 (= p_{22})	0.508 (= $p_{2.}$)
Total	0.439 (= $p_{.1}$)	0.560 (= $p_{.2}$)	1.

$$\text{Odds ratio} = \frac{(n_{11})(n_{22})}{(n_{12})(n_{21})} = \frac{\frac{p_{11}}{p_{12}}}{\frac{p_{21}}{p_{22}}}.$$

$$\text{Odds ratio} = \frac{(151)(192)}{(140)(109)} = \frac{0.255/0.236}{0.184/0.324}$$

$$\text{Odds ratio} = 1.9.$$

d. SAS FREQ and CATMOD code for odds ratio.

```
*READ DATA FOR PRECIPITATION
*HALO, AND WEIGHT
*
*;
DATA HALO;
INPUT PRECIP HALO WT @@;
CARDS;
1     1     151     1     0     109
2     1     140     2     0     192
;
PROC FORMAT;
VALUE PRECIP 1 = ‘YES’ 2 = ‘NO’;
VALUE HALO   1 = ‘YES’ 0 = ‘NO’;
PROC FREQ DATA = HALO     ORDER = DATA;
WEIGHT WT;
TABLE PRECIP * HALO/MEASURES NOROW NOCOL NOPERCENT;
FORMAT PRECIP PRECIP. HALO HALO.;
PROC CATMOD DATA = HALO;
WEIGHT WT;
DIRECT HALO;
MODEL PRECIP = HALO/WLS;
FORMAT PRECIP PRECIP. HALO HALO.;
RUN;
```

e. Output of SAS FREQ.

```
Estimates of Relative Risk    (Row 1/Row 2)
 Type of Study                Value
Case Control (odds ratio)     1.9
```

f. Output of SAS CATMOD

Analysis of Weighted-Least-Squares Estimates

Effect	Parameter	Estimate
Intercept	1	-0.5661
Halo	2	0.6418

Odds ratio estimate = $e^{0.6418}$ = 1.9

Chapter 8

BEST GUESS INTERPRETATION OF MEAN, MEDIAN, AND MODE

8.1 Introduction. Statisticians are frequently asked to identify a single statistic that provides the "best" measure of a typical or average value for a given meteorological variable. Statisticians refer to these as measures of central tendency. The three most common measures of central tendency are the mean, mode, and median. This chapter presents definitions of each of these measures, along with descriptions of some of their advantages and disadvantages.

8.2 Discussion.

a. The mean is the sum of the observations, divided by the number of individual observations. The median (for ungrouped data) is the value of the middle observation, when all the observations are arranged in either ascending or descending order. The mode is the observation that occurs most frequently in the data set. These measures of central tendency are considered typical (or average) in the sense that they are sometimes used to represent all the individual observations of a given element. As an example, consider the following data set: {23, 25, 25, 26, 28, 31, 33, 33, 37, 37, 37, 37, 41}.

1) The sum of all the observations is 413, and there are 13 observations. The mean is then given by (413/13) = 31.8.

2) The median is the value in the middle of the sorted data. With 13 observations, the median is the 7th value, which is 33 in this case (half of the values are less than the median, half are greater.)

3) The value 37 occurs four times. This is more frequent than any other value, so the mode is 37.

b. There is really no way to say, in general, which is the best measure of central tendency. Each has its strengths and weaknesses. The best choice depends upon what you are trying to summarize about the data distribution.

c. Suppose you wanted to guess the value of an observation picked at random from a data set, and you wanted to be correct the highest percentage of the time. In this case, the best guess is the mode.

1) The mode is the value which occurs most frequently. Thus, in the case of a random sample, it is the most likely to be selected.

2) In general, however, the mode is inferior to the mean and median as a measure of central tendency. Often data sets have more than one mode (multimodal distributions). Cloud-cover distributions often have two modes (bimodal) - completely clear and completely overcast. Even when a distribution has only one mode (unimodal), it is possible for the mode to be at an extreme value rather than at a typical value. Consider daily snowfall amount data, the mode at Scott AFB is a trace, but does the mode describe the data set sufficiently?

d. The median is the best guess of a typical value if one wants to come as close as possible to the average, regardless of the sign of the error (absolute error).

1) In descriptive statistics, the median is frequently used. (In descriptive statistics, one merely describes what has occurred — compare this with inferential statistics described in paragraph e. 1, page 20).

2) The median appears to be a better measure of central tendency than the mean when dealing with extreme observations. The mean is strongly affected by large numbers, even when they have small probabilities.

3) Unless you are dealing with a data set that follows (or nearly follows) a normal (Gaussian) distribution, the median, together with quartiles, gives a more precise representation of the distribution than does the mean and standard deviation (Brooks and Carruthers 1953). The median is the value of the 50th percentile, the quartiles are the values of the 25th percentile and the 75th percentile.

e. The mean provides the best guess for any randomly selected variable, provided that one is interested in making the signed error (deviation from the mean) as small as possible (zero, on the average). The mean is also a best guess if one wishes to make the average squared error as small as possible.

1) In inferential statistics, the median is usually inferior to the mean (In *inferential statistics*, one uses what has been observed, the random sample, to infer properties of the population.). The median is difficult to work with mathematically, while the mean is very easy to work with.

2) In some cases, the mean can be anything but a typical value. Consider this extreme case: a station has the following monthly cloud-cover distribution: clear—75 percent, scattered—5 percent, broken—5 percent, overcast—15 percent. The mean cloud-cover is 0.16. But a cloud cover of 0.16 cannot be considered typical when the whole range of values from 0.01 to 0.49 occurs only 5 percent of the time. In fact, cloud-cover distributions often have modes at the extreme values (clear and cloudy), and as a result the mean is often the *least likely* value to occur.

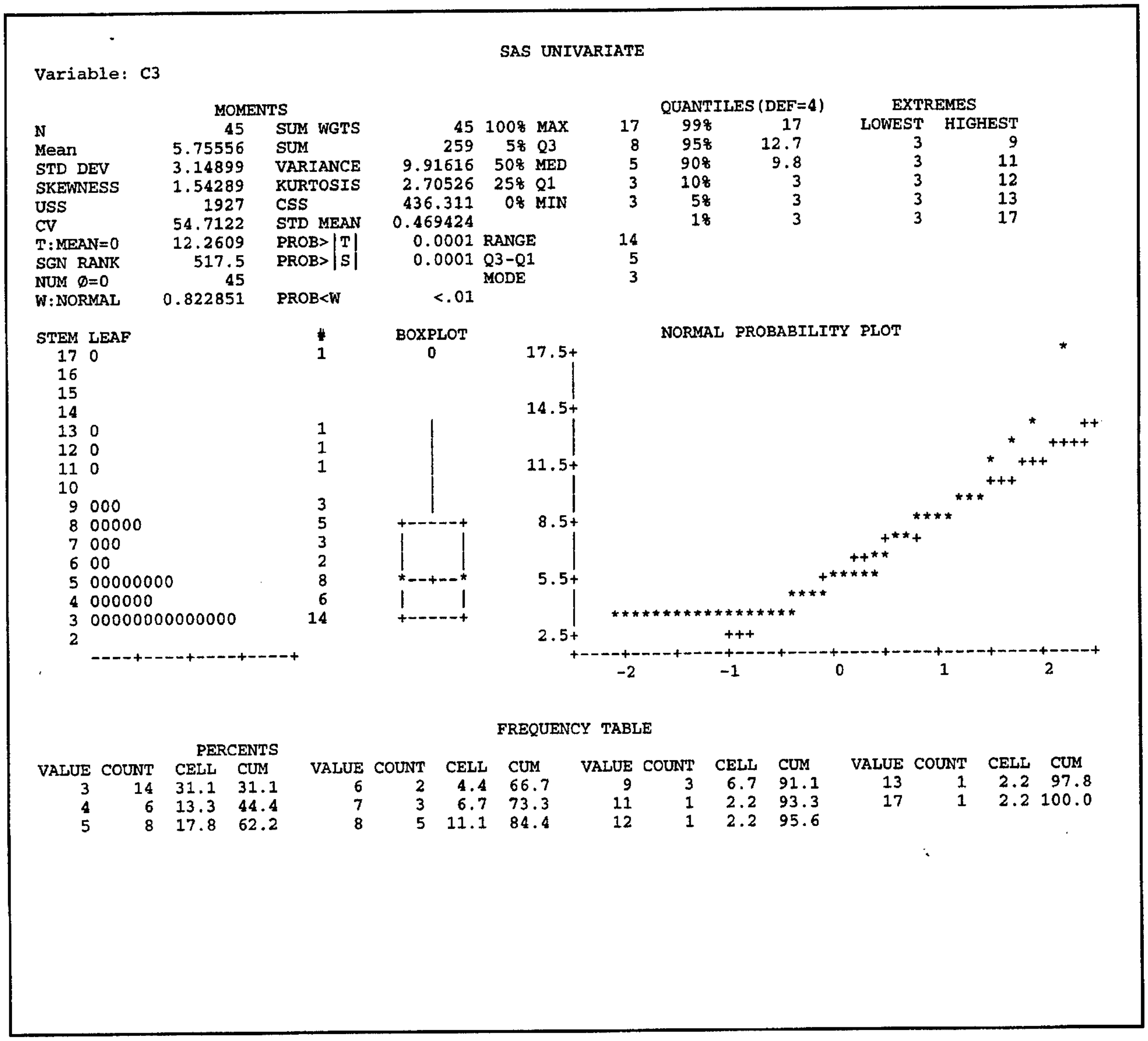
8.3 Conclusion. The choice of a measure of central tendency depends on the distribution of the data, and on what one is trying to infer or describe about the data distribution. It is often useful to compare the mean, median, and mode. When all three are about the same, you can be confident this value can be considered a typical (or average value). When the three values are very different, careful analysis of the data is required to select the best guess of the "typical" value. For symmetrical data, the mean, mode, and median are all equal.

Chapter 9

INTERPRETATION OF SAS UNIVARIATE OUTPUT

9.1 Introduction. The SAS UNIVARIATE procedure produces a wide range of descriptive statistics. This section provides explanations of some of the results. Table 9-1 is an example of a typical SAS UNIVARIATE output.

Table 9-1. Example of output from PROC UNIVARIATE.



```
                                            SAS UNIVARIATE
Variable: C3

                 MOMENTS                                     QUANTILES(DEF=4)     EXTREMES
N                  45   SUM WGTS         45 100% MAX    17    99%       17     LOWEST  HIGHEST
Mean          5.75556   SUM             259   5% Q3      8    95%     12.7          3        9
STD DEV       3.14899   VARIANCE    9.91616  50% MED     5    90%      9.8          3       11
SKEWNESS      1.54289   KURTOSIS    2.70526  25% Q1      3    10%        3          3       12
USS              1927   CSS         436.311   0% MIN     3     5%        3          3       13
CV            54.7122   STD MEAN   0.469424                    1%        3          3       17
T:MEAN=0      12.2609   PROB>|T|     0.0001 RANGE       14
SGN RANK        517.5   PROB>|S|     0.0001 Q3-Q1        5
NUM Ø=0            45                       MODE         3
W:NORMAL     0.822851   PROB<W         <.01

STEM LEAF                  #       BOXPLOT                  NORMAL PROBABILITY PLOT
  17 0                     1          0        17.5+                                              *
  16                                               |
  15                                               |
  14                                           14.5+
  13 0                     1          |            |                                           *    ++
  12 0                     1          |            |                                         *   ++++
  11 0                     1          |        11.5+                                       *  +++
  10                                  |            |                                       +++
   9 000                   3          |            |                                    ***
   8 00000                 5       +-----+      8.5+                                ****
   7 000                   3       |     |         |                             +***+
   6 00                    2       |     |         |                          ++**
   5 00000000              8       *--+--*      5.5+                       +*****
   4 000000                6       |     |         |                    ****
   3 00000000000000       14       +-----+         |   ******************
   2                                            2.5+              +++
     ----+----+----+----+                          +----+----+----+----+----+----+----+----+----+----+
                                                        -2        -1         0         1         2

                                            FREQUENCY TABLE
               PERCENTS
VALUE COUNT  CELL  CUM    VALUE COUNT  CELL  CUM    VALUE COUNT  CELL  CUM    VALUE COUNT  CELL  CUM
    3    14  31.1  31.1       6     2   4.4  66.7       9     3   6.7  91.1      13     1   2.2  97.8
    4     6  13.3  44.4       7     3   6.7  73.3      11     1   2.2  93.3      17     1   2.2 100.0
    5     8  17.8  62.2       8     5  11.1  84.4      12     1   2.2  95.6
```

9.2 Discussion.

a. ***Variable.*** The name of the variable analyzed in the UNIVARIATE procedure ("C3" in this example).

b. *N.* The number of observations. In this example N = 45.

c. ***Mean.*** The "average" value. The mean is defined as the sum of all the values of the variable, divided by the total number of observations.

$$\text{mean} = (\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{259}{45} = 5. \quad \textbf{(1)}$$

d. ***Standard Deviation.*** The standard deviation is defined as the square root of the variance. This gives a measure of the dispersion of all values around the mean. The formula for estimating the standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\text{corrected sum of squares}}{n-1}} \quad \textbf{(2)}$$

$$s = \sqrt{\frac{436.311}{44}} = 3.15.$$

When we have a symmetrical bell-shaped distribution, about 68 percent of the cases in the sample will fall between the limits of ±1 standard deviation from the mean. About 95 percent will lie between ±2 standard deviations, and nearly all the cases (99.75 percent) between ±3 standard deviations.

e. ***Skewness.*** Skewness is a measure of how nonsymmetric a distribution is. If the data is normally distributed, then the computed skewness value will be close to zero. Positive skewness indicates that the distribution has a long tail to the right (mean > median > mode). A distribution is negatively skewed if the left tail is longer (mode > median > mean). See Figures 9-1a-c.

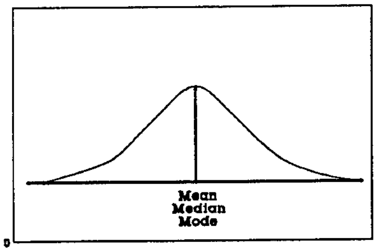


Figure 9-1a. Symmetrical.

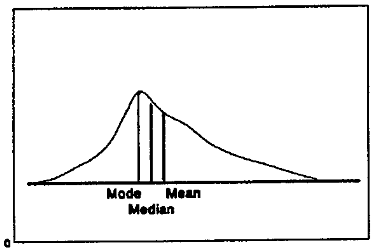


Figure 9-1b. Positively skewed.

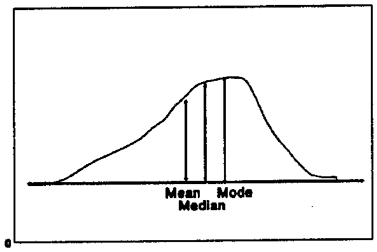


Figure 9-1c. Negatively skewed.

f. ***USS***. Uncorrected sum of squares, denoted by Sx_i^2. The uncorrected sum of squares can be used to compute the variance (s^2) as follows:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n - 1} . \qquad \textbf{(3)}$$

g. ***CV***. The coefficient of variation (CV) is a unitless statistic used to compare the dispersion of two or more distributions. The coefficient of variation is obtained by dividing the standard deviation by the mean. SAS expresses the coefficient of variability as a percentage, but does not include the percent sign. For variable C3 a CV of 54.7122 is 54.7122 percent. To better understand this term, consider the following hypothetical example for annual rainfall rates:

Tropical Station:

Sample Mean = 100 inches
Standard Deviation = 18 inches

Midlatitude Station:

Sample Mean = 30 inches
Standard Deviation = 8 inches

At a glance, rainfall rates at the tropical station may appear to vary more than rainfall rates at the midlatitude station. However, rainfall rates at the tropical station vary less as a percentage of their own mean than do those at the midlatitude station.

Coefficient of variation at tropical station: $\overline{x}$ = 0.18 or 18 percent.

Coefficient of variation at midlatitude station: $\overline{x}$ = 0.26 or 26 percent.

Therefore, the variability for the rainfall rates of the midlatitude station is greater than the variability for the rainfall rates of the tropical station. The midlatitude station's standard deviation is 26 percent of its mean, while the tropical station's standard deviation is 18 percent of its mean.

h. ***T: Mean = 0***. Student's **t** value for testing the hypothesis that the population mean is zero.

i. ***Prob > ½T½*** If this probability value is less than 0.05, then we can say that according to Student's **t** value the population mean is not equal to zero, at the five percent level of significance.

j. ***Sgn Rank***. The signed rank test is a substitute for Student's t-test. It is a nonparametric test; there is no requirement for the data to be normally distributed.

k. ***Num*** $\neq$ ***0***. The number of nonzero observations. The number of nonzero observations is 45 for variable C3.

l. ***Prob > ½S½*** If the probability value is less than 0.05, then we can say that the population mean is not equal to zero at the five percent level of significance, according to the signed rank test.

m. ***W: Normal***. Statistical test for normality. The Shapiro-Wilks **W** statistic is performed when the number of observations is less than 2,000. The Kolmogorov-Smirnov **D** statistic is calculated for samples larger than 2,000.

n. ***Prob < W*** *or* ***Prob > D***. Probability value for testing the hypothesis that the data comes from a normal distribution. If the probability value is less than 0.05, then the conclusion of either the Shapiro-Wilks W statistic or Kolmogorov-Smirnov D Statistic is that we dot not have a normal distribution at the five percent level of significance.

o. ***Sum Wgts***. The sum of the weights of the observations. For unweighted data, the sum of the weights is identical to the number of observations.

p. ***Sum***. The total of the observations. For variable C3 the sum is 250. Use the sum to compute the mean.

q. ***Variance***. The variance is a measure of the distribution of all values around the mean value. The formula for finding the variance, s^2, is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1} \qquad \textbf{(4)}$$

where x_i is the *i*th data point and $\bar{x}$ is the mean. Variance for variable C3 is

$$\frac{\text{corrected sum of squares}}{n - 1} = \frac{436.31}{44} = 9.92. \quad \textbf{(5)}$$

By knowing the mean and the variance of a distribution, one can estimate the probability that a given value, or range of values, will be observed.

r. ***Kurtosis***. Kurtosis is a measure of "tail heaviness." For a normal distribution, kurtosis will be zero (according to the SAS formula). Some authors define a coefficient as 3.0 for a normal distribution, however, they omit the subtraction of 3.0 in their kurtosis formula. A distribution with a high peak, steeper than the normal distribution, exhibits positive kurtosis. A distribution that is flatter than the normal will have a negative value of kurtosis. Kurtosis will be negative for a bimodal curve. See Figure 9-2.

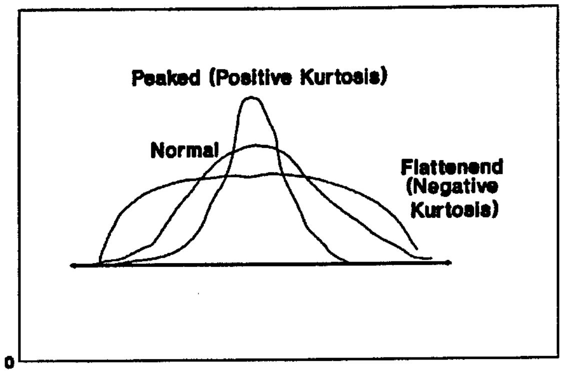


Figure 9-2. Kurtosis

s. ***CSS***. Corrected sum of squares ($\sum(x_i - \bar{x})^2$). The corrected sum of squares is used to compute the standard deviation and variance as discussed in paragraphs 2d and 2q.

t. ***Std Mean***. The standard error of the mean. Standard error of the mean is the standard deviation (3.14899) divided by the square root of n (45), the number of observations. The standard error of the mean is used to obtain the student's **t** value.

$$\text{Std Mean} = \frac{\text{Std Dev}}{\sqrt{n}} = \frac{3.15}{\sqrt{45}} = 0.47. \quad \textbf{(6)}$$

$$t = \frac{5.76}{0.47} = 12.26.$$

u. ***Quantile***. The term quantile is not as well known as the terms median, quartile, decile, and percentile. The median is the 0.5 quantile, the upper and lower quartiles are the 0.75 and 0.25 quantiles, and the 99 percentile is the 0.99 quantile.

v. ***100 percent Max***. The maximum value. The maximum value for variable C3 is 17.

w. ***Q_1, Median, Q_3***. There are three values (Q_1, median, Q_3) that partition a frequency distribution into four equal parts:

1) ***Q_3***. Q_3 is the value at the end of the third quarter of the sequence of measured values, ordered by size. For variable C3 a Q_3 of 8 tells the analyst that 8 is greater than 75 percent of the 45 values.

2) ***Median***. The median is that value in the sequence of individual values, ordered according to size, which divides the sequence in half. It is important to note that the median is not influenced by extreme values, whereas the arithmetic mean is rather sensitive to extreme values. A median of 5.0 for variable C3 tells the analyst that 50 percent of the 45 values are above 5.0, and 50 percent of the values are below 5.0.

3) ***Q_1***. Q_1 is the value that lies at the end of the first quarter of the sequence of measured values, ordered by size. For variable C_3 a Q_1 of 3 tells the analyst that 3 is greater than 25 percent of the 45 values.

x. ***0 percent Min***. The smallest value. The smallest value for variable C3 is 3.

y. ***Range***. The range is the difference between the highest and lowest values, and measures the spread of the data. A range of 14 represents a difference between 17 (highest) and 3 (lowest).

z. ***Q_3 - Q_1***. The difference between the upper and lower quartiles. For variable C3 the difference between Q_3 (75 percent quartile) and Q_1 (25 percent quartile) is 8 - 3 = 5.

aa. ***Mode***. The most frequent sample value. For symmetrical (unimodal) distributions, the mean, median, and mode are equal. For variable C3 the most frequent sample value is 3 with a count of 14.

bb. ***99 percent, 95 percent, 90 percent, 10 percent, 5 percent, 1 percent Percentiles***. The 99th, 95th, 90th, 10th, 5th, and 1st percentile values. The *p*th percentile is defined as a point below which *p* percent of the cases fall. A 99th percentile of 17 means that 99 percent of the data for variable C3 falls below 17. (Percentiles are sometimes called deciles.)

cc. ***Extremes***. The five largest and five smallest values.

dd. ***Stem Leaf.*** A stem-and-leaf plot is printed if N (the number of observations) is no more than 48. A horizontal bar chart is printed if the number of observations is greater than 48. A stem-and-leaf display is an adaptation of a histogram. In a stem-and-leaf display, the bars are proportional to the number of data points in each class. The entries allow the analyst to see how the data are distributed within each such class.

ee. ***Box plot.*** A box plot is used to summarize a set of data in terms of a few easily obtained and understood numbers. The bottom and top edges of the box are located at the sample 25th and 75th percentiles. The center horizontal line is drawn at the sample median and the central plus sign (+) is at the sample mean. An interquartile range is the distance between the 25th and 75th sample percentiles. Any value more extreme than this is marked with a zero if it is within three interquartile ranges of the box, or with an asterisk (*) if it is still more extreme.

ff. ***Normal Probability Plot.*** If the data was from a normal distribution, the normal probability plot should approximate a straight line. Asterisks (*) mark the data values. The plus signs (+) provide a reference straight line. If the data was from a normal distribution, they should tend to fall along the reference line. A large number of visible + signs indicate a non-normal distribution. If the sample is from a normal distribution, then the asterisks form a straight line and this covers most of the + signs.

gg. ***Frequency Table.***

1) ***Value.*** Frequency table of variable values. Variable C3 ranged from 3 through 17.

2) ***Count.*** For variable C3, the value 3 occurred 14 times in a sample of 45.

3) ***Cell.*** For variable C3 a value of 3 occurred 14/45 or 31.1 percent of the time in a sample size of 45.

4) ***Cumulative.*** A cumulative frequency distribution shows, for each cell, the total number of observations in all cells up to and including that cell.

Chapter 10

T TESTS FOR COMPARISONS OF INDEPENDENT SAMPLES AND PAIRED DATA

10.1 Introduction. The t test is a statistical tool that can be used to determine if values from two data sets are statistically different from each other. One of two variations of the t test is used, based on whether the two samples being tested consist of independent data, or paired (correlated) data. If the samples are independent, the "standard" t test is used. The "paired" t test is used if the two samples are correlated. This chapter describes the SAS procedures used for these t tests.

10.2 T Test for Independent Samples.

a. The SAS procedure PROC TTEST is used to compare two independent samples. PROC TTEST tests the hypothesis that the difference between the two sample means is zero.

b. An example (taken from Brooks and Carruthers) will help to illustrate the use of this t test. Suppose we want to determine if the frequency of gale strength winds during the years 1894-1903 is the same as the frequency from 1912-1914. The variable GALE represents the frequency of gales, TIME represents the year, a "0" is used to identify data taken from 1894-1903, and a "1" represents data from 1912-1914.

c. The null hypothesis being tested is that the two population means are equal (the mean gale frequency from 1894-1903 equals the mean gale frequency from 1912-1914). We test this hypothesis with the t test formula for independent samples, assumption of equality of population variances, and unequal sample sizes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \quad . \qquad (1)$$

"Pooled" standard deviation is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \, . \qquad (2)$$

In this example the value for t is -4.21. The calculated t value is compared to a critical value from a statistical table, based on the appropriate degrees of freedom and level of significance (in this case the degrees of freedom is $n_1 + n_2 - 2 = 10$). The t value for 10 degrees of freedom, at a level of significance of a = .05 is 2.201. Since the calculated t is greater than the critical value, one rejects the null hypothesis and assumes the population means are not equal.

d. The following SAS code is used to run this test:

```
DATA;
INPUT GALE TIME;
CARDS;
24      0
14      0
7       0
3       0
5       0
2       0
9       0
5       0
6       0
27      1
21      1
36      1
;
PROC TTEST;
CLASS TIME;
VAR GALE;
RUN;
```

This SAS program will generate the following output:

```
                                   TTEST PROCEDURE

    Variable:  Gale
                          Std       Std
   Time     N    Mean     Dev      Error     Min      Max
   ----     -    ----     ---      -----     ---      ---
    0       9    8.33     6.85      2.28     2.0     24.0
    1       3    28.0     7.55      4.36    21.0     36.0

   Variances      T       DF         Prob >|T|
   ---------      -       --         ---------
   Unequal     -3.996    3.2          0.0258
   Equal       -4.214    10.0         0.0018

   For Ho: Variances are equal, F' = 1.21    DF = (2,8)    PROB > F' = 0.6934.
```

e. The SAS TTEST procedure performs two tests, one which assumes equal variances, and one that does not. Consult the results of the F test at the bottom of the printout to decide which t test to use. In this case the "PROB > F" value is greater than our level of significance (0.6934 > 0.05), so one can accept the null hypothesis that the population variances are equal. Based on this conclusion, the t test results on the second line of output (assuming equal variances) are used. Here, the "PROB > T" value is less than our level of significance (0.0018 < 0.05), so the null hypothesis that the population means are equal is rejected.

10-3. T Test for Paired Observations.

a. In cases when observations are not independent of each other, use of the t test described above is inappropriate. The SAS procedure PROC MEANS is used to compare two samples consisting of paired observations (data sets that are not independent of each other). In this case the hypothesis being tested is slightly different from the previous example. PROC MEANS uses a paired t test to test the hypothesis that the mean of the differences between the two samples is equal to zero.

b. To use PROC MEANS, you must first create a new variable (DIFF) containing the differences between the paired variables. The options "T" (t test) and PRT" (probability value associated with t) within PROC MEANS can then be used to test whether the mean difference is different from zero.

c. The following example will help illustrate the use of the t test with paired data. In this example, X and Y represent the paired observations (for example, temperature measurements taken at the same place and time by two different instruments). The data is shown in Table 10-1.

The formula for the paired t test is

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{3}$$

where $\bar{d}$ = the mean of the differences
and s_d = the standard deviation of the difference

$$s_d = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^2 - \frac{1}{n}(\sum_{i=1}^{n} d_i)^2\right]}. \tag{4}$$

In this case t = 2.798. From a table of t values, the critical t value is 2.365 (at a 95 percent confidence level, with 7 degrees of freedom). Since the t value is greater than the critical value, analysts reject the null hypothesis and conclude that the mean difference is not equal to zero.

Table 10-1. Example paired data.

Obs.	X_i	Y_i	D_i	D_i^2
1	4.0	3.0	1.0	1.00
2	3.5	3.0	0.5	0.25
3	4.1	3.8	0.3	0.09
4	5.5	2.1	3.4	11.56
5	4.6	4.9	-0.3	0.09
6	6.0	5.3	0.7	0.49
7	5.1	3.1	2.0	4.00
8	4.3	2.7	1.6	2.56

d. The SAS code to run this test is shown below.

```
DATA;
INPUT X Y;
DIFF = X - Y;
CARDS;
(data)
;
PROC MEANS MEAN STDERR T PRT;
VAR DIFF;
RUN;
```

This SAS program will generate the following output:

```
Analysis Variable:  DIFF

 Mean    STD ERROR      T      PROB > |T|
 1.15      0.411      2.798    0.0266
```

Here, the “PROB > |T|” value is less than the significance level (0.0266 < 0.05), so the null hypothesis is rejected and it's concluded that the mean difference is not equal to zero.

Chapter 11

SKEWNESS AND KURTOSIS

11.1 Introduction. The SAS UNIVARIATE procedure calculates skewness and kurtosis statistics, which are used for detecting deviations from normality. However, the SAS formulas for calculating skewness and kurtosis are different from the formulas used in most statistics textbooks. This chapter defines the skewness and kurtosis statistics, and relates the SAS formulas to other skewness and kurtosis formulas.

11.2 Discussion.

a. ***Skewness*** is a measure of the non-symmetry of a distribution (see Figures 11-1a through 11-1c). Skewness values can be positive or negative. Positive skewness indicates the data distribution has a larger "tail" to the right (mean > median > mode). A distribution is negatively skewed if the left tail is larger (mode > median > mean).

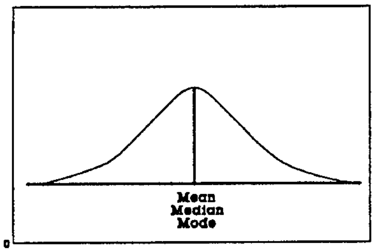


Figure 11-1a. Symmetrical.

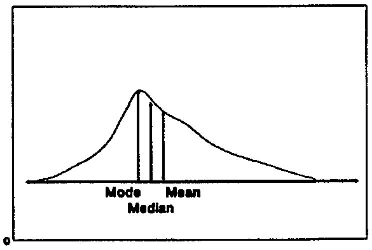


Figure 11-1b. Positively skewed.

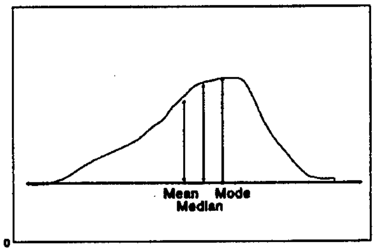


Figure 11-1c. Negatively skewed.

b. ***Kurtosis*** values discriminate between a "peaked" and a "flat-topped" distribution (see Figure 11-2). Kurtosis values can be either positive or negative. Peaked distributions often have positive kurtosis values, indicating there is a large number of data values near the mean. A negative kurtosis value is often associated with a flat-topped distribution.

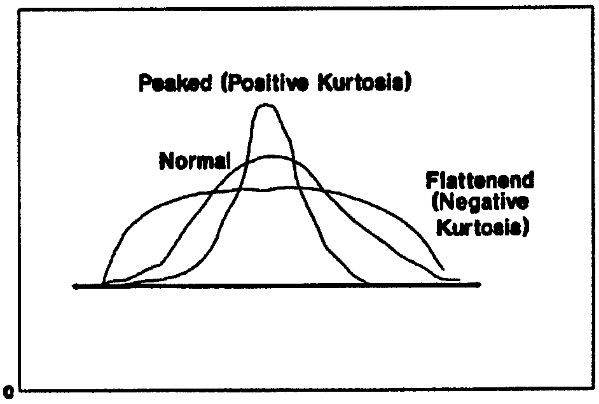


Figure 11-2. Kurtosis.

c. The formulas commonly used in statistics textbooks to compute skewness and kurtosis are shown below. Skewness and kurtosis are normally defined in terms of "moments." The mathematical notations used are $\sqrt{b_1}$ (skewness) and b_2 (kurtosis).

(1) Skewness formula:

$$\sqrt{b_1} = \frac{m_3}{m_2\sqrt{m_2}} \qquad \textbf{(1)}$$

where:

$$m_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n} \qquad \textbf{(2)}$$

$$m_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad \textbf{(3)}$$

(2) Kurtosis formula:

$$b2 = \frac{m_4}{m_2^2} \qquad \textbf{(4)}$$

where

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}. \qquad \textbf{(5)}$$

d. As mentioned previously, the formulas used by SAS to calculate skewness and kurtosis differ from those commonly used in textbooks. The SAS UNIVARIATE procedure does not generate the $\sqrt{b_1}$ (skewness) and b_2 (kurtosis) statistics. Instead, the Fisher g statistics are calculated: g_1 (skewness) and g_2 (kurtosis). The g_1 and g_2 statistics are related to the $\sqrt{b_1}$ and b_2 statistics by the following procedure:

$$\sqrt{b_1} = \frac{(n - 2)}{\sqrt{n(n - 1)}} g_1 \qquad \textbf{(6)}$$

$$b_2 = \frac{(n - 2)(n - 3)}{(n + 1)(n - 1)} g_2 + \frac{3(n - 1)}{n + 1} \qquad \textbf{(7)}$$

where n is the sample size, g_1 is the SAS skewness value and g_2 is the SAS kurtosis value.

e. Interpretation of skewness, $\sqrt{b_1}$, and kurtosis, $b_{2,}$ values according to Chou (1975). When $|\text{skewness}| \geq 1$, the distribution is highly skewed. When $0.5 < |\text{skewness}| < 1$, the distribution is moderately skewed. When $0 < |\text{skewness}| < 0.5$, the distribution is nearly symmetric. When the b_2 statistic, known as kurtosis, is equal to 3, the data is said to be normal (bell shaped). When $b_2 < 3$, the distribution is said to be platykurtic (flat). When $b_2 > 3$, the distribution is said to be leptokurtic (peaked). Chou's interpretation of b_2 can be further evaluated against the baseline of 3, or b_2 -3 to obtain positive and negative kurtosis values. If $b_2 < 3$, then b_2 -3 results in a negative kurtosis value. If $b_2 = 3$, then b_2 - 3 results in a kurtosis value of 0. If $b_2 > 3$, then b_2 -3 results in a positive kurtosis value.

f. The SAS UNVIARIATE procedure does not evaluate the skewness and kurtosis statistics, but we can use a "rule of thumb" to derive meaning from the values calculated. If the SAS UNIVARIATE procedure generates skewness and kurtosis values close to 0, then we can assume that the data is normally or symmetrically distributed. If a more detailed interpretation is desired, the technique taken from Snedecor (1980) can be used as shown below:

(1) Skewness is confirmed and the distribution is non-normal if

$$\left|\sqrt{b_1}\right| > 4\sqrt{\frac{6}{n}}. \qquad \textbf{(8)}$$

(2) Kurtosis is confirmed and the distribution is non-normal if

$$|b_2 - 3| > 4\sqrt{\frac{24}{n}}. \qquad \textbf{(9)}$$

Chapter 12

L-MOMENT METHOD AND DISCORDANCY MEASURE

12.1. Introduction. This chapter gives definitions and formulas for L-moments, definition and formulation of the discordancy measure, and use of the discordancy measure statistic to quality control or flag the data.

12.2. Discussion.

a. The theory behind L-moments is not that new, but Hosking (1990) developed a unified approach that has become a competitor to other statistical techniques with the today's computers and statistical software.

b. Certain linear combinations of the ranked observations (ordered from smallest to largest) in a random sample are called L-moments. Observations in a random sample must be ranked to compute L-moments. L-moments (l_1, l_2, l_3, l_4) and L-moment ratios r_3 and r_4 are useful for summarizing distributions. The first L-moment (l_1) is the arithmetic mean , while the second L-moment (l_2) is a measure of dispersion similar in certain aspects to the standard deviation. The L-moment ratio r_3 or l_3/l_2 is referred to as L-skewness and is a measure of symmetry. The L-moment ratio r_4 or l_4/l_2 is referred to as L-kurtosis and is a measure of peakedness. The third and fourth L-moments (l_3, and l_4) are used to formulate the measures of skewness and kurtosis, respectively. The L-coefficient of variation or l_2/l_1 is a measure of spread relative to the size of the numbers in a sample.

c. Hosking (1990) outlines L-moment formulas for random samples of probability distributions when there is knowledge of the cumulative distribution function (e.g., Normal or Weibull). A specific cumulative distribution may be specified by its L-moments. Hosking also outlines L-moment formulas for random samples from an unknown distribution. This chapter addresses only the formulas for random samples from an unknown population because L-moments are usually estimated from a random sample drawn from an unknown population.

d. Formulas of L-moments estimated from a random sample drawn from an unknown distribution. Consider the random sample $x_1, x_2, \ldots, x_n$ ordered from smallest to largest as $x_{1:n} \leq x_{2:n} \leq \ldots \leq x_{n:n}$ (an ordered random sample).

$$l_1 \text{ (first L-moment)} = \frac{\sum_{i=1}^{n} x_i}{n} .$$

$$l_2 \text{ (second L-moment)} = \frac{1}{2}\binom{n}{2}^{-1} \sum_{i>}\sum_{j} (x_{i:n} - x_{j:n}).$$

$$l_3 \text{ (third L-moment)} = \frac{1}{3}\binom{n}{3}^{-1} \sum_{i>}\sum_{j>}\sum_{k} (x_{i:n} - 2x_{j:n} + x_{k:n}).$$

$$l_4 \text{ (fourth L-moment)} = \frac{1}{4}\binom{n}{4}^{-1} \sum_{i>}\sum_{j>}\sum_{k>}\sum_{l} (x_{i:n} - 3x_{j:n} + 3x_{k:n} - x_{l:n}).$$

$$L\text{-}skewness\ (r_3) = \frac{l_3}{l_2} .$$

$$L\text{-}kurtosis\ (r_4) = \frac{l_4}{l_2} .$$

$$L\text{-}coefficient\ of\ variation = \frac{l_2}{l_1} .$$

Note the second L-moment may also be equivalently written as

$$l_2 = \binom{n}{2}^{-1} \frac{1}{2} \sum_{i=1}^{n} (2i - n - 1) x_i$$

a computationally easier formulation. It follows that simpler variations of the third and fourth L-moments may also exist, and the reader may find it worthwhile to investigate the literature further before attempting calculations using the formulas given above.

e. In the literature, conventional moments are usually used to estimate skewness and kurtosis. Conventional moments are the average of different powers of a random variable. Conventional moments denoted by the average values of third and fourth powers about the mean are associated with measures of skewness and kurtosis. According to Hosking (1990), the main advantages of L-moments or linear combinations of data over conventional moments are the smaller impact of outliers and the more confident inferences derived from smaller samples. Hosking feels that conventional moments result in biased results because conventional moments require squaring and cubing the observations which causes them to give greater weight to the larger observations. A disadvantage of the conventional skewness and kurtosis formulas is that when skewness and kurtosis values are calculated from finite samples, the sample skewness and kurtosis values are bounded and it is unusual for the sample skewness and kurtosis to attain the full range of values available to the population skewness and kurtosis. In contrast, the sample L-moment skewness and kurtosis values calculated from a sample of size $n \geq 4$ can take any of the feasible values of the population L-moment skewness and kurtosis.

f. Discordancy Measure.

1) Background. Hosking and Wallis (1993) have proposed a discordancy measure to flag data that needs to be checked. The discordancy measure can be used to identify those sites that are grossly discordant with the group or cluster as a whole. Discordancy is measured in terms of the L-moments of the sample data. Hosking and Wallis (1993) provide the framework for the discordancy measure. Discordancy measure is a guideline rather than a formal statistical test.

2) Formal Definition. Let

$$u_i = [t_1^i \;\; t_2^i \;\; t_3^i \;]^T$$

be a vector containing L-coefficients of variation, L-skewness, and L-kurtosis for individual site i. Let T denote transpose of row vector. Then let

$$\bar{u} \text{ (mean)} = \frac{\sum_{i=1}^{n} u_i}{n}$$

be the mean of all the initial sites i. Define the sample covariance matrix as

$$S = (n - 1)^{-1} \sum_{i=1}^{n} (u_i - \bar{u})(u_i - \bar{u})^T .$$

S^{-1} represents the inverse of matrix S. Define the discordancy measure for site i as

$$D_i = \frac{1}{3} (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}) .$$

3) Interpretation of discordancy measure. Large values of D_i indicate sites that are most discordant from the group as a whole. A given site is declared discordant if $D_i > 3$. This measure allows the analyst to identify those sites whose L-moments are not consistent with other sites in a group and should be moved to other groups.

g. Example of data quality check. Vogel and Lin (1992) has an alternative quality control measure that is simpler to calculate than the discordancy measure. The alternative measure is also a quality control check of data using L-skewness and L-kurtosis. If individual L-skewness and L-kurtosis of an individual site is greater than three standard deviations from the mean L-skewness and mean L-kurtosis of individual skewness and kurtosis values for a group of sites, then the individual site is flagged and needs to be checked.

12.3 Summary.

a. L-moments can only be applied to random samples.

b. L-moments are more difficult to calculate than conventional moments; however, Hosking (1991) has FORTRAN routines for using the method of L-moments.

c. According to Vogel, the L-moment strategy is being used by the National Weather Service in some applications. However, other authors in the field are not that impressed with L-moments. L-moments have a tendency to give conservative values, and it is difficult to determine if more conservative values are better or not in some cases. Most people continue to use the conventional moment formulas in their analysis to determine skewness and kurtosis.

Chapter 13

CALCULATING PERCENTILES USING SAS UNIVARIATE PROCEDURE

13.1 Introduction. Percentiles can provide useful measures of variability. By definition, the p^{th} percentile of a set of measurements arranged in order of magnitude is that value that has at most percent of the measurements below it, and at most (100 - p) percent above it. This chapter describes how to calculate various default and non-default percentiles using the SAS UNIVARIATE procedure.

13.2 Discussion.

a. Percentiles divide a sample into 100 parts. The term percentile is frequently associated with similar statistical terms. For example, deciles (which divide a sample into ten parts) are often called the 10th, 20th, . . .90th percentile. Quartiles (which divide the data into four parts) are termed the 25th , 50th, and 75th percentile.

b. The SAS UNIVARIATE procedure automatically calculates the 1st, 5th, 10th, 90th, 95th, and 99th percentile. Any other percentile from 0 to 100 may also be calculated using the PCTLPTS option. The examples below show various ways of calculating percentiles with PROC UNIVARIATE.

13.3 Example. The following data set, consisting of 40 ordered data points, will be used in our examples:

```
2 2 2 3 3 3 4 4 4 4
4 5 5 5 5 5 6 6 6 6
6 6 7 7 7 7 7 7 7 7
8 8 8 8 9 9 9 9 10 10
```

a. Percentile formula. Once the observations in a data set have been ordered by magnitude, a given percentile can be calculated as follows: first calculate the quantity (np/100) + 1, where n is the number of observations, and p is the percentile of interest. If the quantity (np/100) is not an integer, then the p^{th} percentile is the sample observation with order number (np/100). If (np/100) is an integer, the p^{th} percentile is the average of the two sample observations with order numbers (np/100) and (np/100) + 1. For example, the 60th percentile of the data set shown above is calculated as follows:

$$n = 40; p = 60$$
$$np/100 = 24.$$

Since this is an integer, the 60th percentile is the average of the 24th and 25th observation. These observations are both 7, so the 60th percentile is (7 + 7)/2, which is 7.

b. SAS UNIVARIATE examples. To calculate any percentile other than the default values provided by PROC UNIVARIATE, the options PCTLPTS and PCTLPRE must be used in the OUTPUT statement. The option PCTLPTS specifies which percentiles you wish to calculate. PCTLPRE specifies prefixes used to create variable names for the percentiles requested (similarly, the option PCTLNAME may also be used to create suffixes).

1). Output of default percentiles. The following SAS code will create a data set called "RESULTS", containing the default percentiles listed.

```
DATA;
INPUT TEMP;
CARDS;
(data)
;
PROC UNIVARIATE;
VAR TEMP;
OUTPUT OUT=RESULTS  P1=P1 P5=P5
P10=P10 P90=P90 P95=P95
P99=P99;
RUN;
```

2). Output of non-default percentile with a percentile name. In this case, the data set "RESULTS" will have the variable C_DEG, containing the 20th percentile value.

```
PROC UNIVARIATE;
VAR TEMP;
OUTPUT OUT=RESULTS  PCTLPTS=20  PCTLPRE=C_
PCTLNAME=DEG;
RUN;
```

(3) Output of non-default percentiles with a default percentile name. In this example the data set "RESULTS" will have the variables C_50, C_95, C_97.5, and C_100, with the corresponding percentiles. The PCTLPTS option serves as a default name when the PCTLNAME option is not used.

```
PROC UNIVARIATE;
VAR TEMP;
OUTPUT OUT=RESULTS  PCTLPTS=50, 95 TO 100 BY 2.5
PCTLPRE=C_;
RUN;
```

Chapter 14

CORRELATION COEFFICIENTS

14.1 Introduction. This chapter describes two methods of measuring the degree of association between variables: the Pearson product-moment correlation, and the Spearman rank coefficient.

14.2 Discussion. The default correlation coefficient used in the SAS procedure PROC CORR is the Pearson product-moment correlation (Pearson's r). Pearson's r shows the degree of the linear relation between two normally distributed variables. If either variable is non-normally distributed, a nonparametric measure of association such as the Spearman rank correlation (Spearman's r) should be used.

a. The Pearson product-moment correlation is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\ .$$

b. The Pearson product-moment correlation coefficient is only a measure of linear relation. It's of no use in describing nonlinear relations between two variables x_i and y_i. A zero value for r does not necessarily imply that no relationship exists between x_i and y_i, only that there is no linear relationship. Positive values of r indicate a tendency for x_i and y_i to increase together. Negative values of r indicate that as x_i increases, y_i tends to decrease, and vice versa. See Figure14-1 to see the relationship between x_i and y_i at various values of Pearson's r.

c. The SAS procedure PROC CORR prints out a probability value below the sample Pearson product-moment correlation coefficient (r). If this probability value is less than 0.05 (95 percent level of confidence) or 0.01 (99 percent level of confidence), then we can say that the population correlation coefficient is not equal to zero. Therefore, some degree of linear relation exists between x_i and y_i.

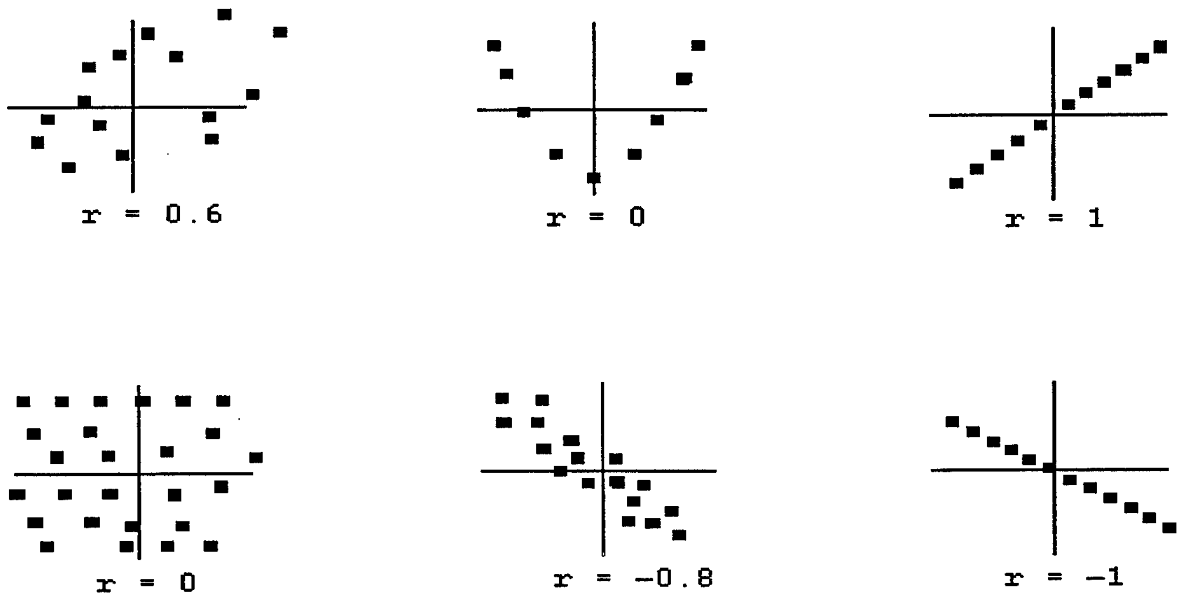


Figure 14-1. Sample scatter plots.

d. The Pearson product-moment correlation is affected by outliers. A method to check if outliers have had an effect on the correlation coefficient is to compare a Pearson product-moment correlation based on the total data set with a Pearson product-moment correlation from a data set where the outliers have been removed.

e. Pearson's r is appropriate only for normally distributed data. When any variable appears substantially non-normal, a nonparametric correlation coefficient such as Spearman's r, should be used. Spearman's r is calculated by first arranging the data in order of increasing or decreasing magnitude. The lowest observation is referred to as rank 1, the next as rank 2, and so forth to the last observation. If two or more observations have the same value, they are given an average rank (EXAMPLE: for the data set {1, 2, 2, 3}, the respective ranks are {1, 2.5, 2.5, 4}). The Spearman rank correlation is then computed using the formula for Pearson's r, on the ranked data. It is best to think of Spearman's r as a measure of association or agreement. The size of the Spearman coefficient does tell something about the tendency of the variables to relate in a monotone-increasing or monotone-decreasing way.

Chapter 15

COVERAGE USING THE BOEHM ALGORITHM AND TETRACHORIC CORRELATION

15.1 Introduction. If the probability of an event occurring at a single point is known, what is the corresponding probability of the same condition occurring along a line of sight, or in the surrounding area? This problem was recently addressed by Al Boehm, during the development of the C_CLOUD_S (Climatology of Cloud Statistics) program. The Coverage Using Boehm (CUB) algorithm that he developed can be used to obtain the probabilities of fractional coverages. This chapter discusses the CUB algorithm, and the required input variables (including the tetrachoric correlation matrix) needed to obtain a probability value.

15.2 Discussion.

a. It's not possible to estimate probabilities for fractions along a line-of-sight, or for portions of an area from historical records. However, the CUB algorithm, which computes the cumulative probability that a given coverage threshold will not be exceeded, can be used for this purpose. Some of the terms used in the application of the CUB algorithm are:

Cover - The desired fraction of the domain (portion of a line or area). Cover ranges from 0 to 1.

Mean Probability - The mean probability refers to a known point in the domain. The mean probability is obtained from historical records, and has a range from 0 to 1.

Mean Correlation - The mean correlation ($\bar{r}$) can be either the mean of all the Pearson product-moment correlations (r_{ij}), or tetrachoric correlations (r_{ij}) between "n" stations in the "n x n" correlation matrix.

$$\bar{r} = \frac{1 + r_{12} + r_{13} + r_{21} + 1 + r_{23} + r_{31} + r_{32} + 1}{n^2}.$$

This paper centers on only the mean of the tetrachoric correlations because Boehm feels that the tetrachoric correlation is more robust than the Pearson product-moment correlation. All n x n correlations between n stations are considered. For example, if there are 3 stations (n=3), then the 3x3 correlation matrix is

$$\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}.$$

Note: Even though $r_{12} = r_{21}$, $r_{13} = r_{31}$, and $r_{23} = r_{32}$, all intercorrelations of this equicorrelated matrix are summed to find the mean correlation ($\bar{r}$).

Degrees of Freedom (DOF) - The degrees of freedom is the number of *independent* dimensions (stations, in this case). The degrees of freedom is calculated as follows:

$$DOF = \frac{1 - 2\bar{r} + \bar{r}^2}{\overline{r^2} - (\bar{r})^2}$$

$$where \ \ \overline{r^2} = \frac{(1)^2 + (r_{12})^2 + (r_{13})^2 + (r_{21})^2 + (1)^2 + (r_{23})^2 + (r_{31})^2 + (r_{32})^2 + (1)^2}{n^2}.$$

b. The CUB algorithm is provided below, followed by a definition of the terms used.

$$CUB = PNORM\left(\frac{-(ENORM(MEAN)\ -\ RMEAN)}{\sqrt{RVAR}}\right).$$

$$RVAR = ROBAR + (1 - ROBAR) \times ESTME\,.$$

$$RMEAN = \sqrt{(1 - ROBAR)} \times EM\,.$$

PNORM - This function computes the inverse transnormalization (described below).

ENORM - This function computes the transnormalization (described below).

MEAN - The historical mean probability.

RMEAN - Represents the equivalent normal deviate of the mean.

RVAR - The variance of the transnormalized mean.

ROBAR - The mean correlation.

ESTME - Interpolated variance between the variance of the median and the variance of the end point (end points are when cover = 0.0, and when cover = 1.0).

EM - The transnormalized mean with respect to the cover and the degrees of freedom

EM = ENORM (0.5 + (COVER - 0.5)($2^{(1 - FOD)}$ - 1)) where FOD is 1 over the degrees of freedom.

c. Definition of transnormalization and inverse transnormalization. Since many meteorological variables are not normally distributed, transformations must be performed before statistical manipulations are conducted. In the CUB algorithm, ENORM is the function used for this purpose. ENORM performs a transnormalization by converting the cumulative probability value of a variable to an equivalent normal deviate (END). The ENDs tend to be normally distributed, with a range from -4.5 to 4.5. The inverse transnormalization function (PNORM) is used to return to the cumulative probability of the variable.

d. Tetrachoric Correlation. The tetrachoric correlation coefficient is useful for estimating the degree of association between two variables (or two stations) for which we have only dichotomized (yes/no) information. Boehm (1993) reports that the tetrachoric correlation is more conservative, that is it varies less at different locations and times, and is more robust against outliers. Boehm uses the tetrachoric correlation formula of Panofsky and Brier (1965) in his computer program, but he treats the tetrachoric correlation formula result as an estimate of the first guess in an iterative solution. Transnormalization and inverse transnormalization are performed between the first guess and the final estimate of the tetrachoric correlation in the iterative solution (Willand, 1992). Using the standard contingency table notation as shown on the next page:

Table 15-1. Contingency table notation.

		STATION A	
		RAIN	NO RAIN
STATION B	RAIN	A	B
	NO RAIN	C	D

The tetrachoric equation is defined as

$$r_t = \sin\left[\frac{\pi}{2}\frac{\sqrt{AD}-\sqrt{BC}}{\sqrt{AD}+\sqrt{BC}}\right].$$

The tetrachoric equation is accurate when (A+B)/N = 0.5, and (A+C)N = 0.5, but for values near zero or one, the equation is in error. For this reason, Boehm uses an iterative solution or linear extrapolation to find a better estimate of the tetrachoric correlation.

15.3 Example. A recent project dealing with precipitation data provides an example of how to apply the CUB algorithm. The problem involved estimating the probabilities of various amounts of precipitation coverage within a given area. The number of rain days for 16 stations within the area was known, making it possible to compute a correlation matrix, a portion of which is shown in Table 15-2. With this input data the CUB algorithm produced the coverage probabilities shown in Table 15-3.

Table 15-2. Tetrachoric correlation matrix.

	SZL	COU	SGF	TOP	UIN
SZL	1	0.87	0.81	0.82	0.78
COU	0.87	1	0.82	0.72	0.84
SGF	0.81	0.82	1	0.68	0.69
TOP	0.82	0.72	0.68	1	0.68
UIN	0.78	0.84	0.69	0.68	1

Table 15-3. Coverage probabilities (cumulative) computed by the CUB algorithm.

Coverage	Probability	Coverage	Probability
0	0.421	0.55	0.753
0.05	0.466	0.6	0.773
0.1	0.506	0.65	0.794
0.15	0.542	0.7	0.813
0.2	0.575	0.75	0.832
0.25	0.605	0.8	0.85
0.3	0.633	0.85	0.868
0.35	0.66	0.9	0.886
0.4	0.685	0.95	0.903
0.45	0.708	1	0.92
0.5	0.731		

Chapter 16

AUTOCORRELATION

16.1 Introduction. Consider a series of observations $\{x_1, x_2, x_3, ..., x_n\}$. If the value of x_i is unaffected by any of the remaining values of x, then $\{x_1, x_2, x_3, ..., x_n\}$ is said to form a series of independent observations. Meteorological observations are not usually independent of preceding conditions, though the dependency decreases with the length of time between successive events. The tendency for the occurrence of a specific event to be more probable at a specified time, given that it has occurred in the preceding time period, is known as persistence. According to Brooks and Carruthers (1953), a measure of persistence is the coefficient of autocorrelation. This chapter will center on the definition, application, and interpretation of autocorrelation.

16.2 Background.

a. The Pearson correlation coefficient (r) refers to the degree of relationship between two variables. It is a measure of how one variable will vary, given changes in the value of the second variable. The Pearson correlation coefficient can vary from -1 (which indicates a perfectly linear negative correlation) to +1 (which indicates a perfectly linear positive correlation). When r is greater than 0, the two variables are positively correlated (as one increases, the second one also tends to increase); when r is less than zero they are negatively correlated (as one increases, the other one tends to decrease).

b. Autocorrelation (also known as serial correlation) has many properties which are similar to the Pearson correlation. Autocorrelation differs in the respect that only one variable is considered at a time. Autocorrelation involves the correlation between different values of a single variable at different time intervals. Autocorrelation shows how a variable relates to itself for a specified time lag (a time lag is the length between time periods). A time lag of 1 implies a difference of one time period. An autocorrelation of one time lag indicates how values of periods {1, 2, 3, 4, ...} correlate with values of periods {2, 3, 4, 5, ...}. An autocorrelation of 12 time lags indicates how values of periods {1, 2, 3, 4, ...} correlate with values of periods {12, 13, 14, 15, ...}.

c. The formula for the Pearson correlation coefficient is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_1 - \bar{y})^2}} \qquad (1)$$

where $\bar{x}$ is the sample mean for variable x, and $\bar{y}$ is the sample mean for variable y.

d. The autocorrelation coefficient, r_k is

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \qquad (2)$$

where k is the number of time lags separating y_t and y_{t+k}.

e. The SAS procedure PROC ARIMA can be used to generate autocorrelation values, as shown below.

```
PROC ARIMA;
IDENTIFY VAR = variable name;
RUN;
```

f. Table 16-1, next page, provides an example of the SAS printout from PROC ARIMA. A positive autocorrelation at lag 1 can be an indication of persistence or serial correlation. An autocorrelation near zero indicates a case of completely random data. It should be noted that autocorrelation estimates will not necessarily be exactly equal to zero when the observations are independent.

Table 16-1. An example of SAS printout from PROC ARIMA.

```
ARIMA Procedure

                                       Name of variable  =  S.

                               Mean of working series  =         6
                               Standard deviation      =  3.464102
                               Number of observations  =        10

                                            Autocorrelations

Lag   Covariance   Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1       Std Err
 0        12.000       1.000     |                   |********************|             0
 1         8.200       0.683     |                   |**************      |         0.316
 2         5.000       0.417     |                   |********            |         0.440
 3         1.900       0.158     |                   |****                |         0.478
 4        -1.100      -0.092     |                 **|                    |         0.483
 5        -3.100      -0.258     |             ******|                    |         0.485
 6        -4.600      -0.383     |           ********|                    |         0.498
 7        -4.900      -0.408     |           ********|                    |         0.527
```

g. The SAS printout provides the standard errors next to the sample autocorrelations for various lags. The standard error of a sample autocorrelation can be used to tell whether the autocorrelation is significantly nonzero. Statistically, a sample autocorrelation is regarded as being significantly different from zero (with roughly 95 percent confidence) if it is larger in magnitude than twice its standard error. For example, the autocorrelation at lag 1 (0.68) in the SAS printout has a standard error of 0.316. Twice this value is 0.632, which is less than the autocorrelation value of 0.68. Hence, one can conclude with 95 percent confidence, that the sample autocorrelation at lag 1 is significant. **Note:** There is a 5 percent chance that any given autocorrelation will appear to be significant when it is not. Law and Kelton (1991) state that good estimates of autocorrelation at lag 1 are difficult to obtain unless the sample size is large.

Chapter 17

CANONICAL CORRELATION

17.1. Introduction. The Experimental Long-Lead Forecast Bulletin is issued quarterly by the Climate Prediction Center of National Weather Service and is intended to present experimental long-lead forecasts such as three month temperature forecasts. The forecasts are mentioned in Climate Outlook, available on Internet. Canonical correlation methods are one of the strategies being used for the three month temperature forecasts issued by the Climate Prediction Center. At the 12th Conference on Probability and Statistics in The Atmospheric Sciences (1992), Chu and He (1992) delivered a paper on the prediction of Hawaiian winter rainfall using canonical correlation. This chapter will focus on the definitions of canonical correlation and how canonical correlation in the SAS CANCORR procedure can be used in a prediction equation.

17.2. Discussion.

a. Multiple Regression predicts a single dependent variable from a set of multiple independent variables while canonical correlation analysis enables us to predict multiple dependent variables from multiple independent variables.

b. Definitions.

Criterion Variables - Dependent variables. In the example, the dependent variables are A, B, and C.

Predictor Variables - Independent variables. In the example, the independent variables are D, E, and F.

Canonical Variable - Canonical variable is referred to as linear composite or linear combination of criterion or predictor variables. Canonical variable for criterion variables is denoted by ATMOS(i) where ATMOS(i) = $u_1A + u_2B + u_3C$. Canonical variable for predictor variables is denoted by RAIN(i) where RAIN(i) = $v_1 D + v_2E + v_3F$.

Canonical Function - Pair of canonical variables, one for the set of criterion variables and one for the set of predictor variables. In our example the canonical function is (ATMOS(i), RAIN(i)).

Number of Canonical Functions - Maximum number of possible canonical functions that can be extracted from the set of variables equals the number of variables in the smallest data set, independent or dependent. In our example, the maximum number of possible canonical functions is three, (ATMOS1, RAIN1), (ATMOS2, RAIN2), and (ATMGSS, RAIN3).

Canonical Correlation - Measures the strength of the overall relationship between the canonical variables. When the canonical correlation is squared, the canonical correlation represents the amount of variance in one canonical variable that is accounted for by the other canonical variable.

Canonical Roots - Squared canonical correlations are referred to as canonical roots or eigenvalues.

Canonical Coefficients (canonical weights) - Canonical weights are the coefficients of the canonical variables. The traditional approach to interpreting canonical functions involves examining the sign and magnitude of the canonical weight assigned to each predictor or criterion variable in computing the canonical function. Variables with larger weights contribute more to the canonical function. Variables whose weights have opposite signs would exhibit an inverse relationship with each other and those with the same sign a direct relationship. It is customary to standardize the canonical coefficients so that each canonical variable has a variance of one. Standardized results are considered when the different variables have different units.

Canonical Loading - Correlation coefficient between the predictor or criterion variables and the desired canonical variables.

Redundancy Index - The amount of variance in one set of observed variables explained by a canonical variable of the other set of variables. The redundancy measure is analogous to multiple regression, R^2. A redundancy index of 28 percent indicates that 28 percent of the variance in the dependent variables has been explained by the canonical variable for the independent variable set. It can be computed for both the criterion and predictor variables.

c. Prediction equation using canonical correlation analysis.

(1) R - Correlation matrix where

$$R = \begin{vmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{vmatrix}.$$

R_{11} - The intercorrelations among the x independent variables.

R_{11}^{-1} - The inverse of R_{11}.

R_{22} - The intercorrelations among the y dependent variables.

R_{22}^{-1} - The inverse of R_{22}.

R_{12} - The intercorrelations of x and y variables.

R_{21} - The transpose of R_{12}.

λ - Canonical correlation matrix.

I MATRIX - An identity matrix which has diagonal elements equal to one while all the other elements are zero.

(2) Given two sets of data (x,y), the SAS CANCORR method provides a maximum correlation between the canonical variables.

(3) Canonical coefficients (canonical weights), canonical variables, and canonical correlations are in a typical SAS output and are generated by solving the following equations:

$$\left(R_{11}^{-1}R_{12}R_{22}^{-1}R_{21} - \lambda^{2}I\right)u_i = 0$$

$$\left(R_{22}^{-1}R_{21}R_{11}^{-1}R_{12} - \lambda^{2}I\right)v_i = 0.$$

(4) A prediction or regression equation can be determined as follows:

$$\hat{y} = (v')^{-1}\lambda u'x - (v')^{-1}\lambda u'\bar{x} + \bar{y}$$

where $\hat{y}$ is predicted value, u and v are the criterion and predictor matrices of all canonical weights (coefficients), u′ and v′ denote the transpose of u and v matrices, $(v')^{-1}$ denotes the inverse matrix, $\bar{x}$ and $\bar{y}$ denote the matrices of all mean criterion and predictor variables, x denotes the criterion variable, and λ denotes the canonical correlation matrix. This equation was documented by Pao-Shin Chu and Yu Xiang He (1992) at the 12th Conference on Probability and Statistics in theAtmospheric Sciences (1992). Chu and He had better results with this equation as opposed to using regression on the raw data.

(5) The four most important types of output information derived through SAS canonical correlation analysis are: (a) the canonical variables, (b) the canonical correlations between the canonical variables, (c) the statistical significance of the canonical correlations, and (d) the redundancy measure of shared variance or the amount of variance in one set of variables explained by a linear composite of the other set of variables. It can be computed for both the dependent and the independent sets of variables.

d. SAS program to calculate canonical correlation.

```
*
*
*PROBLEM: CAN CHANGES OF SUMMER
*CONDITIONS IN THE ATMOSPHERE
*BE USED TO PREDICT CHANGES OF RAINFALL IN WINTER.
*LET A,B,C (ATMOS) BE VARIABLES FOR
*SUMMER CONDITIONS IN THE
*ATMOSPHERE AND D,E,F (RAIN) BE
*VARIABLES FOR RAINFALL IN
*WINTER
*;
DATA ONE;
INPUT A B C D E F;
CARDS;
{DATA}
RUN;
DATA TWO;
PROC CANCORR DATA = ONE ALL;
VAR A B C;
WITH D E F;
RUN;
```

e. Output of SAS CANCORR procedure.

```
                              Squared
         Canonical            Canonical
         Correlation          Correlation          Pr > F
1        0.795608             0.632992             0.0635
2        0.200556             0.040223             0.9491
3        0.072570             0.00566              0.7748

Standardized Canonical Coefficients for the ATMOS (Dependent Variable)

         ATMOS1               ATMOS2               ATMOS3
A        -0.7754              -1.8844              -0.1910
B        1.5793               -1.1806              0.5060
C        -0.0591              -.2311               1.0508

Standardized Canonical Coefficients for the Rain (Independent Variable)

         RAIN1                RAIN2                RAIN3
D        -0.3495              -0.3755              -1.2966
E        -1.0540              0.1235               1.2368
F        0.7164               1.0622               -0.4188
```

Standardized Variance of the ATMOS (Dependent Variable) Canonical Redundancy Analysis

	Their Own Canonical Variables	The Opposite Canonical Variables
	Proportion	Proportion
1	0.4508	0.2854
2	0.2470	0.0099
3	0.3022	0.0016

Standardized Variance of the Rain (Independent Variable) Canonical Redundancy Analysis

	Their Own Canonical Variables	The Opposite Canonical Variables
	Proportion	Proportion
1	0.4081	0.2584
2	0.4345	0.0175
3	0.1574	0.0008

f. Explanation of SAS Output.

(1) Three canonical functions are printed out, (ATMOS1, RAIN1), (ATMOS2, RAIN2), and (ATMOS3, RAIN3). ATMOS1, ATMOS2, and ATMOS3 are canonical variables for dependent variables (A,B,C) and RAIN1, RAIN2, and RAIN3 are canonical variables for independent variables (D,E,F). In the output the maximum number of canonical functions to be initially considered is three because three is the number of variables in the smallest data set, three independent or three dependent variables.

(2) The canonical correlation indicates the strength of the relationship between pairs of canonical variables. When squared, the canonical correlation represents the amount of variance in one canonical variable that is accounted for by the other canonical variable. The most common practice is to analyze only those canonical functions whose canonical correlation coefficients are statistically significant beyond some level, typically .10 or less (Pr > F).

(3) Even though the canonical correlation for canonical function one is significant, further analysis involving redundancy index must be undertaken to determine the amount of the dependent variable variance that is shared with the independent variables.

(4) The redundancy index for the first pair of canonical variables (canonical function one) indicates that 28.54 percent of the variance in the dependent variables has been explained by the canonical variable for the independent variable set. The redundancy index also indicates that 25.84 percent of the variance in the independent variables has been explained by the canonical variable for the dependent variable set.

(5) The redundancy index is the essential evaluator in the SAS canonical correlation output. If the redundancy index is acceptable, then look at the canonical weights and canonical loadings. The question arises as to what is the minimum acceptable redundancy index to justify the interpretation of canonical functions? No generally accepted guidelines have been established. The analyst would have to judge each canonical function in light of its practical significance to the research problem being investigated. A test for the significance of the redundancy index has been developed, although it has not been widely utilized. Canonical loadings have not

been printed out, but they enable us to know the correlation between the observed predictor criterion variable and the canonical variables. Canonical weights or canonical coefficients have been printed out. The canonical weights (sign and magnitude of the number assigned to each variable) are used in computing the linear combinations for canonical functions.

(6) The redundancy index of the SAS output enables us to know if we will have a good predictor or regression equation. SAS CANCORR procedure provides multiple regression analysis options to aid in interpreting the canonical correlation analysis. You can examine the linear regression for each criterion variable on the opposite set of predictor variables.

Chapter 18

COEFFICIENT OF DETERMINATION

18.1 Introduction. In statistics, the term coefficient of determination is frequently used and often misapplied. This chapter presents a basic definition of this coefficient, and discusses some of the problems associated with its use.

18.2 Background.

a. The major uses of regression analysis are for estimation of parameters and means, and for prediction of new observations. The basic linear regression equation is:

$$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + b_0 \quad \textbf{(1)}$$

where b_0, b_1, ..., b_n are the regression parameters; x_1, x_2, ..., x_n are the predictor variables, and y is the predictand. The coefficient of determination (r^2 or R^2), the square of the correlation coefficient, is usually interpreted as measure of the goodness of fit of a regression line to a set of data (It is common to use the notation r^2 when referring to only two variables, and R^2 when more than two variables are involved.). This value can be computed from

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} \quad (2)$$

or

$$R^2 = 1 - \frac{\text{Error Sum of Squares}}{\text{Total Sum of Squares}}. \quad (3)$$

If the explained variation is equal to the total variation, $R^2 = 1$ and there is a perfect fit (all the observations fall on the regression line). When $R^2 = 0$, there is no linear association between the predictors and the predictand variables. Figures 18-1 and 18-2 graphically depict these two extreme cases. In practice, R^2 is not likely to be exactly 0 or 1, but will lie somewhere in between. The closer its value to one, the greater the degree of linear association between the x_i predictors and the y predictand.

b. The value of R^2 tends to be affected by the spacing of the x observations. Values of R^2 will tend to be higher when the x observations in a sample are highly spaced. R^2 can also be made large by including a large number of independent variables in the model. It has therefore been suggested that a modified measure be used, which adjusts for the number of independent variables in the regression model. This measure, referred to as the *a*djusted coefficient of multiple determination (see equation below), may actually become smaller when another predictor is introduced into the regression equation. The unadjusted R^2 can never decrease.

$$\text{Adjusted } R^2 = 1 - (1 - R^2)(\frac{n - 1}{n - p}) \quad (4)$$

where:

n = sample size

p = number of parameters in model.

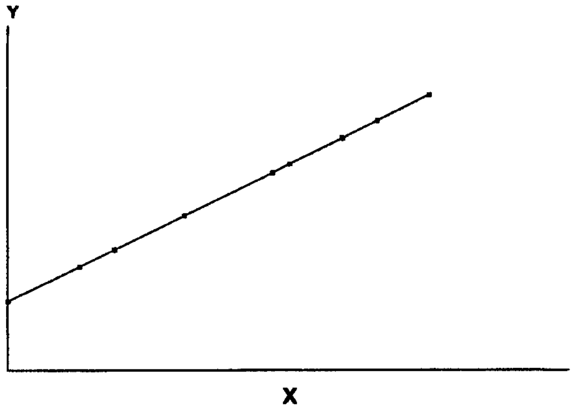


Figure 18-1. Example of regression line for the case when $r^2 = 1$.

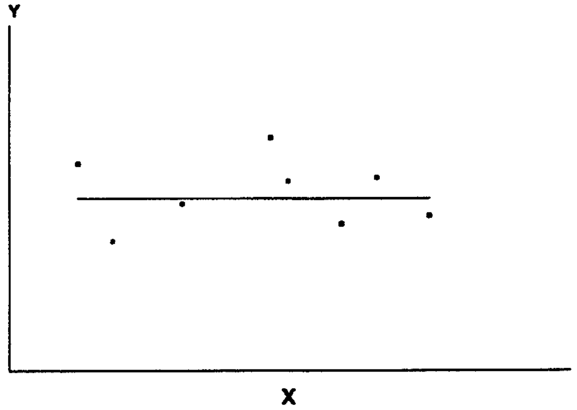


Figure 18-2. Example of regression line for the case when $r^2 = 0$.

c. In regression equations containing multiple variables, multicollinearity may be a problem. Multicollinearity refers to the case in which two or more variables in the regression model are highly correlated, thereby making it difficult to isolate the individual effects on the dependent variable. (An example might be a model that uses both sea level pressure and altimeter settings as independent variables). With multicollinearity, the estimated regression coefficients may be statistically insignificant (and even have the wrong sign!) even though the R^2 is high. The SAS REG procedure has a few techniques available to test for multicollinearity. One way is to examine the variance inflation factor (VIF). If the highest VIF value is larger than 10, one can conclude that multicollinearity is a problem.

18.3 Conclusion. No single measure is an adequate indication of the usefulness of a regression equation. The R^2 measure is simply a descriptive measure of the degree of linear association between x_i and y variables in the sample of observations. However, problems associated with multicollinearity and a large number of predictor variables may cause an analyst to overstate the value of a particular regression model.

Chapter 19

R^2 FOR PREDICTION

19.1 Introduction. In Regression Analysis (SAS REG procedure), the coefficient of determination (R^2) is a measure of the fit of the regression line. The R^2 for prediction is a different measure than R^2 but is often compared with R^2. This chapter will focus on definitions of R^2, R^2 for prediction, and how to interpret the R^2 for prediction.

19.2 Discussion.

a. Definition of R^2 and R^2 for prediction.

(1) R^2 represents the proportion of variation in the response data that is explained by the regression model. The R^2 can range from 0 to 1. The upper bound of R^2 or 1 is achieved when the fit of the model to the data is perfect; i.e., all residuals (differences between observed values and predicted values) are zero.

(2) R^2 for prediction can be compared to R^2 as an aid in determining whether overly influential observations are present. Therefore, the extent to which R^2 for prediction falls below R^2 provides a rough indication of the presence of influential observations.

b. Formula for R^2:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SSTY}$$

where:

$\hat{y}_i$ = predicted response

y_i = observed response

$\bar{y}$ = mean of observed responses

SSE = Error Sum of Squares

SSTY = Total Sum of Squares.

c. Formula for R^2 for prediction:

$$R^2 \textit{ for prediction} = 1 - \frac{PRESS}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where:

$$PRESS(\textit{prediction sums of squares}) = \sum_{i=1}^{n}\left(\frac{e_i}{1 - h_i}\right)^2$$

e_i are the errors and:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

d. Data Example.

OBS	X	Y	Predicted Y Value	Error	Error^2	PRESS Residuals Squared
1	1	5	5.6	-0.6	0.36	4.000
2	2	8	7.2	0.8	0.64	1.306
3	3	9	8.8	0.2	0.04	0.081
4	4	10	10.4	-0.4	0.16	1.778
					1.20 SSE	7.165 PRESS Statistic

e. SAS program to calculate R^2 and R^2 for prediction.

```
*
* READ IN DATA WHERE Y IS
* VARIABLE WE WISH TO
* PREDICT (THE DEPENDENT
* VARIABLE) AND X IS THE
* PREDICTOR VARIABLE (INDEPENDENT
* VARIABLE).
*;
DATA ONE;
INPUT X Y;
CARDS;
  1  5
  2  8
  3  9
  4  10
 RUN;
*
* SAS REGRESSION PROCEDURE
*;
DATA TWO;
PROC REG DATA=ONE;
MODEL Y=X /R NOPRINT;
OUTPUT OUT=TWO PRESS=PRESS
RESIDUAL=ERR;
RUN;
*
* CREATE ERROR SQUARE VARIABLE AND
PRESS VARIABLE
*;
DATA THREE;
SET TWO;
ERRSQ = ERR*ERR;
PRESS=PRESS*PRESS;
RUN;
*
*SAS UNIVARIATE PROCEDURE
*OUTPUTS PRESS STATISTIC,
*SSTY, AND SSE.
*;
DATA FOUR;
SET THREE;
PROC UNIVARIATE DATA=THREE NOPRINT;
VAR Y ERRSQ PRESS;
OUTPUT OUT = FOUR
CSS=SSTY SUM=SUMY SSE PRESS;
RUN;
*
* RSQPRED AND RSQ VARIABLES
*
*;
DATA FIVE;
SET FOUR;
RSQPRED = 1 - PRESS/ SSTY;
RSQ     = 1 - SSE/ SSTY;
RUN;
```

f. Output.

R^2 (RSQ) = 0.914
R^2 for prediction (RSQPRED) = 0.488

19.3 Summary. The value of R^2 for prediction, 0.488, is less than the R^2 value, 0.914. The extent to which R^2 for prediction falls below R^2 provides a rough indication of the presence of influential observations. A comparison of R^2 for prediction with R^2 yields the same information as a comparison of the PRESS statistic with the error sum of squares, SSE. The sum of squared errors, SSE, and the sum of the PRESS residuals squared can be compared as an aid in determining if influential observations are present. The difference between the PRESS residuals squared and the squared error is largest when influential observations are present (see observations 1 and 4).

Chapter 20

LINEAR REGRESSION

20.1 Introduction. The SAS REG procedure builds and analyzes multiple regression equations, producing a wide range of descriptive statistics. This section provides explanations of some of the statistics available in PROC REG using several examples, beginning with the sample output shown in Table 20-1.

20.2 Discussion. Multiple regression equations take the form shown below.

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

where "y" is the dependent response variable, the x's are the independent variables used as predictors, b_0 is the y-intercept, and the other b's are the regression coefficients. Table 20-1 shows the results of a regression equation used to relate "growth" (the dependent variable) to a "dose" of a given drug (DOSE), and the square of the dose (DOSESQ). Using the parameter estimates shown in Table 20-1, the regression model for this example is:

GROWTH = 35.657 + 5.263 (DOSE) -0.128 (DOSESQ)

Example 1. Below is the SAS routine used to produce the output in Tables 1a and 1b.

```
DATA DATA1;
INPUT GROWTH DOSE;
DOSESQ=DOSE**2;
CARDS;

(data)
;
PROC REG DATA = DATA1;
MODEL GROWTH = DOSE DOSESQ /
P R CLI CLM;
RUN;
```

In this SAS routine, "P," "R," "CLI," and "CLM" are four of the many options available in PROC REG. The "R" option requests an analysis of the residuals. The standard errors of the predicted values will be printed, along with the residuals. The studentized residual, which is the residual divided by its standard error, is both printed and plotted. The "CLI" and "CLM" options request 95 percent prediction and confidence intervals for individual predicted values, and for predicted mean values.

The statistics generated by the code shown in example one, as well as other statistics generated when different options are selected, are described more fully in the following text and examples.

a. Dependent (or response) variable. This is the variable we try to predict using the regression equation. In example one the response variable is GROWTH.

b. F value, Prob>F. These entries show the results of the F test used to check the hypothesis that all parameters are zero, a test of the overall significance of the regression. If the "Prob>F" value is greater than 0.05, then the hypothesis is probably true and the regression model is therefore of little value.

c. Root MSE. This is an estimate of the standard deviation of the error term, also known as the standard error of the estimate. As the predictions become more precise, this value will become smaller.

d. Dep. Mean. The sample mean of the dependent value.

e. C.V. The coefficient of variation. This expresses the variation in unitless values (computed as 100 times the Root MSE, divided by the mean of the dependent variable). As a sample of observations becomes more stationary, the C.V. approaches zero.

f. R^2. The coefficient of determination. R^2 measures the fit of a regression line to the sample data (R^2 has values between zero and one). In example one, the R^2 value of 0.9364 indicates that 93.64 percent of the variation in the dependent variable can be associated with the variation in the independent variables used.

g. Adjusted R^2 (ADJRSQ). The R^2 value will always increase as additional predictor variables are added to a regression equation, even when the added predictors have no relation to the response variable. The adjusted R^2 value eliminates this bias. The adjusted R^2 will not automatically increase when new predictors are introduced, and may even decrease if the predictors do not improve the model (in fact, it may become negative). The adjusted R^2 will always be smaller than R^2, but the difference will be negligible if we use a large sample.

h. Variables. These are the variables used to predict the dependent variable. In example one, the predictor variables are DOSE and DOSESQ, along with the y-intercept.

i. Parameter Estimate. The parameter estimates show the value of the y-intercept and the regression coefficients. From the parameter estimates, it is possible to write the regression equation, as shown in Paragraph 2 above.

j. Standard Error. The estimate of the standard deviation of the parameter estimate. The standard error for prediction is actually the standard error for the estimated mean, rather than the standard error of a single predicted value. An "OUTPUT" statement must be used to get both a standard error of the estimated mean and of a single predicted value.

k. t test (T for HO: Parameter = 0), Prob > |T|. These two columns show the results of a t-test used to check the hypothesis that the regression coefficient is equal to zero. If the "Prob > |T|" value for this test is greater than 0.05, then the predictor variable being tested adds little to the model and should be removed.

l. The large box in the bottom half of the SAS output in Table 1 shows the predicted values at each data point, the upper and lower 95 percent confidence and prediction intervals for the mean and predicted values, and information on the residuals (actual minus predicted values).

m. Student Residual. These are obtained by dividing the residuals by the standard errors, with the result following the student's t distribution. For large samples, student residuals larger than 2.5 are rare, thus it is possible to identify unusually large residuals.

n. Cook's D. This statistic is useful in identifying outliers that may have a significant impact on the least squares regression. Cook's D provides an influence measure by calculating the change to the estimate that results from deleting each observation. A Cook's D value greater than one indicates an outlying influential observation.

o. Sum of Residuals. The sum of the residuals should be zero. If a different value is obtained, round off errors are most likely responsible.

p. Sum of Squared Residuals. Excluding round off errors, this should equal the Error Sum of Squares from the top box.

q. Predicted Residual Sum of Squares (PRESS). The PRESS statistic is useful for detecting the existence of outliers. If the PRESS is considerably larger than the residual sum of squares, an influential outlier may be present.

Example 2. The code below describes how to invoke two other options which are available in the REG procedure.

```
PROC REG;
MODEL  Y = X1 X2  X3  X4 / TOL VIF;
RUN;
```

u. VIF and TOL. The variance inflation factor (VIF) is useful in determining which predictor variables may be involved in multicollinearities (high intercorrelation among variables). The tolerance (TOL) is the reciprocal of the VIF. If multicollinearity is present the result can be instability of the regression coefficients and inflated confidence intervals around predicted values. Murphy and Katz (1985) recommend that the use of ridge regression be considered if the maximum VIF is greater than ten. Other statisticians recommend variable deletion be used to reduce multicollinearity.

v. Mallow's Cp. This option, available with the RSQUARE model selection method, helps determine if the model is overfit or underfit. We should look for a Cp value that is about equal to the number of parameters in the regression model.

Example 3.

```
PROC REG;
MODEL Y = X1 X2 X3 X4 /
SELECTION = RSQUARE Cp;
RUN;
```

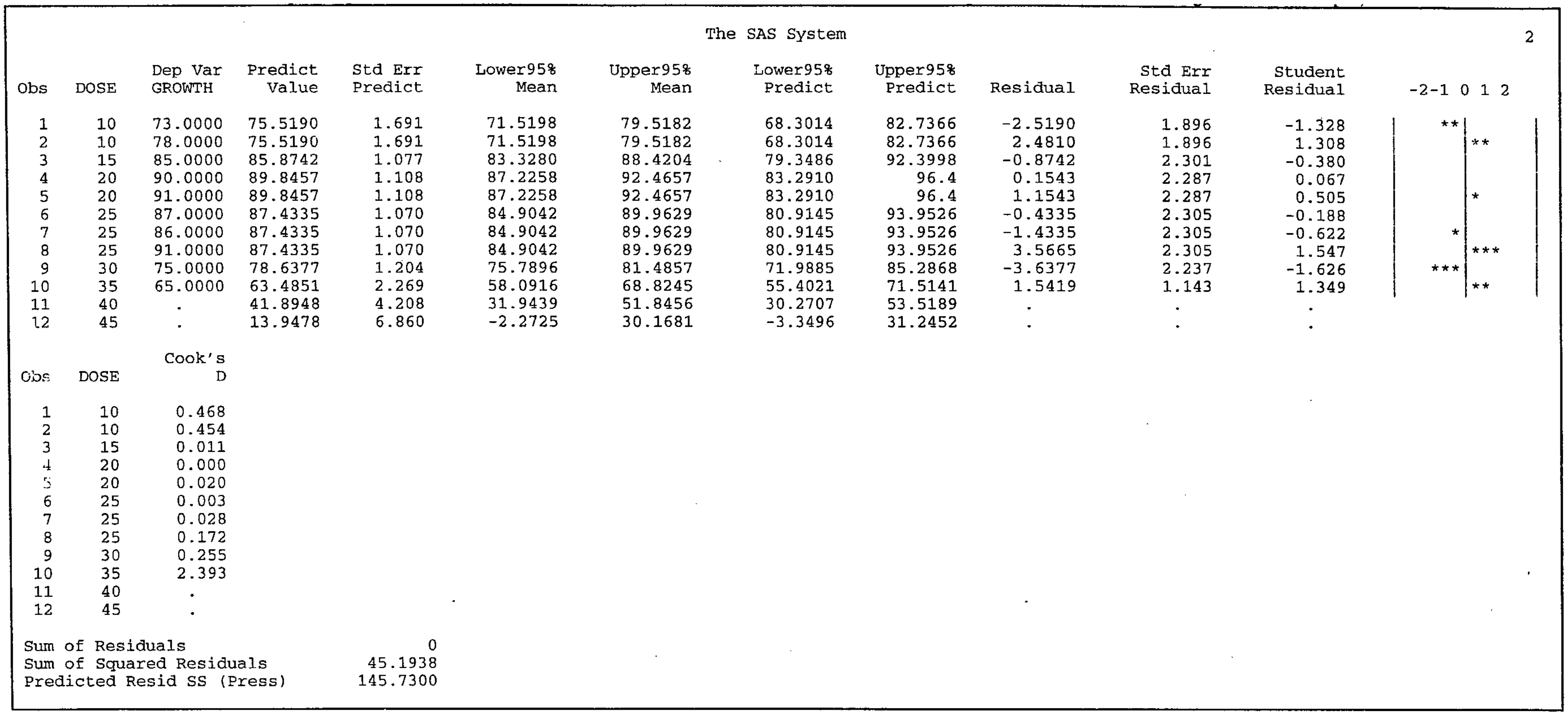
w. INFLUENCE and DW are two additional options that may be useful. The INFLUENCE option enables analysts to flag the influence of each observation on the regression model. Five statistics will be generated (hat matrix, RSTUDENT, COVRATIO, DFFITS, and DFBETAS). If all five statistics exceed a general cutoff value for an observation, then that observation should be investigated to determine why it was found to be so influential. The DW option (Durbin-Watson statistic) can be used to test for serial correlation in the residuals. In general, one can assume that no serial correlation is present in the data if the DW statistic is between 1.5 and 2.5. If there is serial correlation then the R^2 value will be erroneous.

20.3 Conclusion. This section presents a description of some of the information available in the SAS REG procedure. Several additional options not shown in example one are also available within the REG procedure. Consult the SAS User's Guide for additional information on these and other options. In general, a good regression equation will have a large adjusted R^2, a small root mean square error, small Mallow's Cp, and no evidence of multicollinearity.

Table 20-1a. Sample output from PROC REG (from *SAS/STAT User's Guide, Volume* 2). See example 1 for the code used to generate this output.

```
                                The SAS System

Model: MODEL1
Dependent Variable: Growth

                              Analysis of Variance

                           Sum of          Mean
      Source        DF     Squares         Square         F Value        Prob>F

      Model          2     665.70617       332.85309      51.555         0.0001
      Error          7      45.19383         6.45626
      Total          9     710.90000

                    Root MSE        2.54092         R-square        0.9364
                    Dep Mean       82.10000         Adj R-aq        0.9183
                    C.V.            3.09491

                              Parameter Estimates

                           Parameter       Standard       T for H0:
      Variable      DF     Estimate        Error          Parameter=0         Prob > |T|

      INTERCEP       1     35.657437       5.61792724       6.347             0.0004
      DOSE           1      5.262896       0.55802206       9.431             0.0001
      DOSESQ         1     -0.127674       0.01281135      -9.966             0.00001
```

Table 20-1b. Continuation of sample output from PROC REG (from SAS/STAT User's Guide, Volume 2). See example 1 for the code used to generate this output.



```
                                                    The SAS System                                                              2

                Dep Var   Predict   Std Err     Lower95%     Upper95%     Lower95%    Upper95%                  Std Err      Student
 Obs   DOSE     GROWTH      Value   Predict         Mean         Mean      Predict     Predict    Residual     Residual     Residual       -2-1 0 1 2

   1     10    73.0000    75.5190     1.691      71.5198      79.5182      68.3014     82.7366     -2.5190        1.896       -1.328     |     **|     |
   2     10    78.0000    75.5190     1.691      71.5198      79.5182      68.3014     82.7366      2.4810        1.896        1.308     |       |**   |
   3     15    85.0000    85.8742     1.077      83.3280      88.4204      79.3486     92.3998     -0.8742        2.301       -0.380     |       |     |
   4     20    90.0000    89.8457     1.108      87.2258      92.4657      83.2910        96.4      0.1543        2.287        0.067     |       |     |
   5     20    91.0000    89.8457     1.108      87.2258      92.4657      83.2910        96.4      1.1543        2.287        0.505     |       |*    |
   6     25    87.0000    87.4335     1.070      84.9042      89.9629      80.9145     93.9526     -0.4335        2.305       -0.188     |       |     |
   7     25    86.0000    87.4335     1.070      84.9042      89.9629      80.9145     93.9526     -1.4335        2.305       -0.622     |      *|     |
   8     25    91.0000    87.4335     1.070      84.9042      89.9629      80.9145     93.9526      3.5665        2.305        1.547     |       |***  |
   9     30    75.0000    78.6377     1.204      75.7896      81.4857      71.9885     85.2868     -3.6377        2.237       -1.626     |    ***|     |
  10     35    65.0000    63.4851     2.269      58.0916      68.8245      55.4021     71.5141      1.5419        1.143        1.349     |       |**   |
  11     40          .    41.8948     4.208      31.9439      51.8456      30.2707     53.5189           .            .            .
  12     45          .    13.9478     6.860      -2.2725      30.1681      -3.3496     31.2452           .            .            .

               Cook's
 Obs   DOSE         D

   1     10     0.468
   2     10     0.454
   3     15     0.011
   4     20     0.000
   5     20     0.020
   6     25     0.003
   7     25     0.028
   8     25     0.172
   9     30     0.255
  10     35     2.393
  11     40         .
  12     45         .

 Sum of Residuals                       0
 Sum of Squared Residuals         45.1938
 Predicted Resid SS (Press)      145.7300
```

Chapter 21

MULTICOLLINEARITY

21.1 Introduction. Multicollinearity exists when two or more independent variables are highly related to each other. The high correlation between independent variables can cause the computed estimates of regression coefficients to be unstable. This causes the results of the regression to be of limited usefulness. A number of meteorological variables are associated, thus the problem of multicollinearity is often encountered.

21.2 Discussion.

a. When independent variables are highly correlated, the estimated regression coefficients tend to vary widely from one sample to the next. As a result, only imprecise information about the individual regression coefficients may be obtained. Under severe multicollinearity, the regression coefficients may be subject to large round-off errors and large sampling variances.

b. Just as intercorrelations between the independent variables tend to make the estimated regression coefficients imprecise (i.e., erratic from sample to sample), so do the coefficients of partial correlation between the dependent variable and each of the independent variables tend to become erratic from sample to sample.

c. While it may be feasible in multiple regression to vary one independent variable and hold the other independent variables constant, it may not be possible in practice to do so for independent variables that are highly correlated. For example, in a regression model for predicting crop yield from amount of rainfall and hours of sunshine, the relation between the two independent variables makes it unrealistic to consider varying one while holding the other constant.

d. Possible indicators of multicollinearity include:

(1). Large changes in the estimated regression coefficients when a variable is added or deleted, or when observations are altered or deleted.

(2). Important independent variables having nonsignificant results on their regression coefficients.

(3). Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical considerations or prior experience.

(4). Large coefficients of correlation between pairs of independent variables in the correlation matrix.

(5). Wide confidence intervals for the regression coefficients representing important independent variables.

e. Remedial measures for multicollinearity:

(1). One of several independent variables may be dropped from the model in order to reduce multicollinearity and thereby reduce the standard errors of the estimated regression coefficients for the independent variables remaining in the model.

(2). In polynomial regression models, expressing the independent variables in the form of deviations from the mean serves to reduce substantially the multicollinearity among the first-order, second-order, and higher-order terms.

(3). Occasionally, it is possible to add some observations which break the pattern of multicollinearity (not always possible).

(4). Ridge regression may be used to remedy multicollinearity by modifying the method of least squares to allow biased estimators of the regression coefficients.

(5). Use regression with principal components where the independent variables are linear combinations of the original independent variables.

(6). Use Bayesian regression where prior information about the regression coefficients is incorporated into the estimation procedure.

21.3 Conclusion. There is no one definite way to deal with multicollinearity; opinions vary among statisticians. The easiest way to deal with it is to eliminate correlated variables.

Chapter 22

RIDGE REGRESSION

22.1 Introduction. Situations occasionally arise when it is necessary to use highly correlated independent variables in a regression equation, leading to problems with multicollinearity (correlation between independent variables). One solution to this problem is to use ridge regression. Ridge regression reduces the correlations among the independent variables, thereby making it possible to obtain more stable estimates of the regression coefficients.

22.2 Discussion.

a. When regression estimators are closely associated, the least squares estimators of the regression coefficients are subject to large standard errors. As a result, these least squares estimators may differ substantially from those obtained in other studies, even though both produce reasonable results. Ridge regression replaces the least squares estimators with biased estimators. Ridge regression tends to bias regression estimators toward zero.

b. The following example illustrates the principle behind ridge regression. With two independent variable vectors (X_1 and X_2) and a dependent variable vector (Y) in a multiple linear regression equation, the ordinary standardized least squares coefficients are computed as shown below.

$$b_1 = \frac{r_{1Y} - r_{12}\, r_{2Y}}{1 - r_{12}^2} \qquad b_2 = \frac{r_{2Y} - r_{12} r_{1Y}}{1 - r_{12}^2}$$

where r_{1Y}, r_{2Y}, and r_{12} are the Pearson product-moment correlations between X_1 and Y, X_2 and Y; and X_1 and X_2, respectively. The ridge estimators equations are:

$$b_1^* = \frac{r_{1Y} - \left(\frac{r_{12}}{(1+k)}\right) r_{2Y}}{1 - \left(\frac{r_{12}}{(1+k)}\right)^2} \left(\frac{1}{(1+k)}\right)$$

$$b_2^* = \frac{r_{2Y} - \left(\frac{r_{12}}{(1+k)}\right) r_{1Y}}{1 - \left(\frac{r_{12}}{(1+k)}\right)^2} \left(\frac{1}{(1+k)}\right).$$

Notes: 1. The value of k, the ridge trace, is usually a number between 0 and 1.
2. When k = 0, you obtain the ordinary least squares estimates.
3. The value of k is obtained subjectively. When the estimators stabilize, the "proper" k value has been determined.

22.3 Conclusion. Ridge regression is used only when highly correlated predictor variables must be used. Ridge regression techniques are controversial. Murphy and Katz state that if the variance inflation factors (VIF) of the predictors are large, then it is appropriate to consider ridge regression in order to minimize the effects of the predictor variable correlations, and develop a set of stable coefficients (Ridge regression is available in the SAS REG procedure, by specifying the ridge option.).

Chapter 23

SAS ORTHOREG PROCEDURE

23.1 Introduction. The SAS REG and GLM procedures can be used to perform multiple linear regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

where y is the predicted response, x_n are the predictor variables, and b_n are the regression parameters. Multiple linear regression is an attempt to fit a straight line to the collected data. In meteorology, analysts often use predictor variables that are correlated in both space and time. When performing multiple linear regression with correlated variables, the user may obtain the following warning note in the SAS printout:

> **Note**: Model is not full rank. Least-square solutions for the parameters are not unique. Some statistics will be misleading. Parameters have been set to 0 since the variable are a linear combination of other variables. B means the parameter estimate is biased.

Version 6 of SAS has a new regression procedure known as PROC ORTHOREG. This procedure can serve as a strategy to eliminate the warning note in the regression analysis. This section explains the reasons for the above warning message, and discusses how to correct the problem using PROC ORTHOREG.

23.2 Background.

a. PROC REG and PROC GLM have an option to calculate the parameters in a least squares linear equation (multiple linear regression). SAS uses matrix algebra to calculate the parameter estimates. In order to calculate the parameters of the multiple linear equation, an inverse of a matrix must be computed. If it is not computationally possible to carry out this matrix inversion, then the warning message stating the model is not full rank will appear. This indicates that there is a linear dependency between the predictor variables. If the inverse of a matrix cannot be carried out in the usual way, then the matrix is termed singular.

b. The warning message indicates a problem known as collinearity. If only two predictors are involved, it is highly unlikely the warning message will appear. If there are more than three predictors, collinearity may be more likely.

c. PROC REG and PROC GLM use a generalized inverse to compute parameter estimates. This may lead to erratic values for the estimates (wide fluctuations in the parameter estimates as predictors are added or removed from the regression equation) when there is a linear dependency between the predictor variables. The ORTHOREG procedure can prevent the occurrence of erratic parameter estimates.

d. In some cases simple round-off errors can lead to collinearity. Consider the following set of equations:

$$2x + y = 5$$
$$2.000001x + 0.999999y = 5.000001$$

There is a unique solution for x and y from this set of equations; however, a high degree of precision is required to obtain it. If precision in the sixth decimal place is lost, the second equation becomes

$$2x + y = 5.$$

Now the set of equations no longer has a unique solution. In this example, collinearity is due to numerical round-off.

23.3 Example.

a. The following example demonstrates the problems that can arise due to collinearity. Suppose we wish to develop a multiple linear regression model to forecast fog with a visibility of less than or equal to 5,000 meters (a yes/no type model), 3 hours from the observation time. Potential predictors from the current observation include: temperature (TEMP), cosine and sine of the hour angle (CHH and SHH), cosine and sine of the wind direction (CWD and SWD), dew point depression (DDEP), wind speed (WSP), dew point (DEW), wind direction (WDIR), ceiling height (CIG), sea level pressure (SLP), altimeter setting (ALT), and visibility (VSBY). The SAS code to conduct this regression is shown on the next page.

```
DATA ONE;
INFILE DATAIN MISSOVER;
INPUT CHH SHH CWD SWD TEMP DDEP WSP DEW WDIR CIG SLP ALT VSBY;
PROC REG;
MODEL VSBY=CHH SHH CWD SWD TEMP DDEP WSP DEW WDIR CIG SLP ALT VSBY;
RUN;
```

A summary of the output from this procedure is displayed below.

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. Parameter as dew has been set to 0 since the dew variable is a linear combination of the other variables. B means the parameter estimate is biased.

```
DEW = -6.94 • INTERCEPT - 1.7E-09 • CHH  + 1.33E-09 • SHH

        - 2.72E-09 • CWD - 1.55E-09 • SWD  + 1.00 • TEMP

        - 1.00 • DDEP + 1.92E-10 • WSP + 1.35E-11 • WDIR

        - 1.30E-13 • CIG + 1.82E-08 • SLP

        - 6.23E-08 • ALT + 2.83E-13 • VSBY


                              PARAMETER ESTIMATES

        VARIABLES             DEGREES              PARAMETER
                              OF FREEDOM           ESTIMATES

        INTERCEPT             B                    4.654
        CHH                   B                    0.014
        SHH                   B                    -0.003
        CWD                   B                    0.009
        SWD                   B                    -0.027
        TEMP                  B                    0.003
        DDEP                  B                    0.006
        WSP                   B                    0.001
        DEW                   0                    0
        WDIR                  B                    0.0001
        CIG                   B                    0.000000493
        SLP                   B                    -0.001642
        ALT                   B                    0.004078
        VSBY                  B                    0.0000444
```

b. Note that only the values B and 0 appear in the "degrees of freedom" column. Any variable which is a linear combination of other predictor variables (such as DEW) is denoted by 0. Other variables are flagged with a B, for "biased".

c. Now we run the same example with the SAS ORTHOREG procedure. The SAS program is shown below.

```
DATA ONE;
INFILE DATAIN MISSOVER;
INPUT CHH SHH CWD SWD TEMP DDEP WSP DEW WDIR CIG SLP ALT VSBY;
PROC ORTHOREG;
MODEL VSBY= CHH SHH CWD SWD TEMP DDEP WSP DEW WDIR CIG
SLP ALT VSBY;
RUN;
```

A summary of the output from this SAS ORTHOREG procedure is shown below.

PARAMETER ESTIMATES

VARIABLES	DEGREES OF FREEDOM	PARAMETER ESTIMATES
INTERCEPT	1	4.654
CHH	1	0.014
SHH	1	-0.003
CWD	1	0.009
SWD	1	-0.027
TEMP	1	0.003
DDEP	1	0.006
WSP	1	0.001
DEW	0	0
WDIR	1	0.0001
CIG	1	0.000000493
SLP	1	-0.001642
ALT	1	0.004078
VSBY	1	0.0000444

d. It turns out that the fog example had only a mild case of collinearity, so the ORTHOREG procedure generated the same parameter estimates as the REG procedure. This won't always be the case.

23.4 Conclusion. There are different degrees of collinearity. In a mild case of collinearity, an easy approach is to simply eliminate the variables identified as a linear combination of other variables (in this case, DEW). In a severe case of collinearity, the SAS ORTHOREG procedure can be used. A comparison can be made between the results of the two procedures to determine the stability of the parameter estimates. The real test of a linear equation is its performance on independent data.

Chapter 24

THE SAS/ETS AUTOREG PROCEDURE

24.1 Introduction. One of the assumptions in linear regression is that the residuals are independent (a residual is the difference between an observed and a predicted value). Since meteorological data are often in time series form, serial correlations (also called autocorrelations) between data points frequently exist. This chapter describes the SAS AUTOREG procedure, which is a regression technique that is appropriate when serial correlations exist within the data.

24.2 Discussion.

a. Serial correlation between data points can be detected by studying the residual plot from a linear regression. Figures 24-1 through 24-3 show examples of how the residual plots may look for cases of positive, negative, or no serial correlation. Positive serial correlation is characterized by unusually large clusters of positive and negative residuals as in Figure 24-1. In positive serial correlation there is more of a smooth pattern in the residuals. In Figure 24-2, negative serial correlation is illustrated by rapid switching between positive and negative residuals. If no serial correlation exists, a more random pattern of residuals will occur as in Figure 24-3. Serial correlation can also be detected by using the Durbin-Watson statistic (provided by PROC AUTOREG). In general, a D-W statistic between 1.5 and 2.5 (one can look at a D-W table for exact values) indicates no significant serial correlation exists. If the value is less than 1.5, positive serial correlation likely exists. If the D-W value is greater than 2.5 there is negative serial correlation. The further the value is from the range 1.5 and 2.5, the more serious the problem.

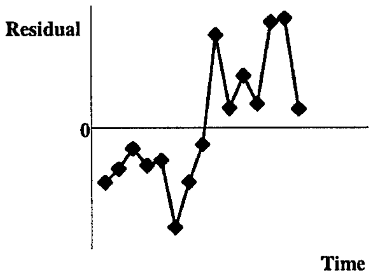


Figure 24-1. Positive Serial Correlation

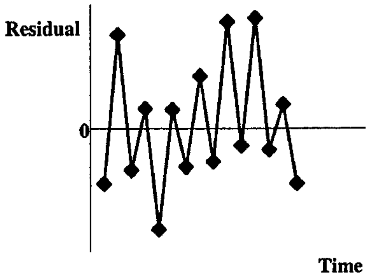


Figure 24-2. Negative Serial Correlation

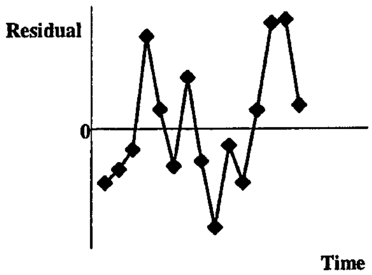


Figure 24-3. No Serial Correlation

b. The AUTOREG procedure corrects for serial correlation by fitting a regression equation to the residual pattern. To see how AUTOREG differs from simple linear regression, let's compare the two. A simple linear regression equation has the form

$$y_i = a + \beta x_i + \varepsilon_i.$$

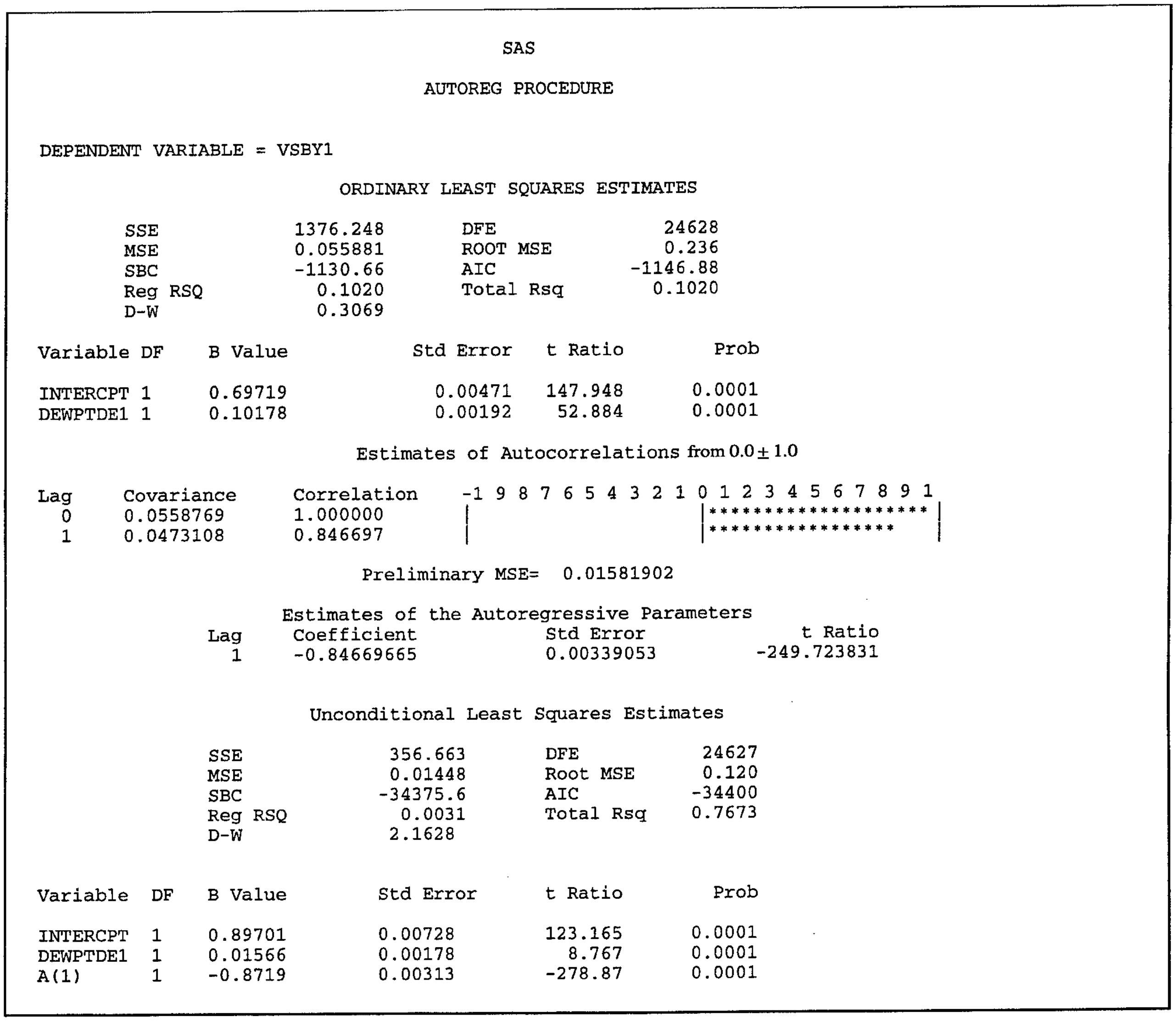
In this equation, e$_i$ represents the model error. It is assumed that the sum of the errors equals zero, and that the errors are normally distributed and independent. These assumptions are often violated. With meteorological data, the error terms are often correlated, as mentioned previously. In this case, the autoregressive model can be used. The equation for this has the form

$$y_t = a + \beta x_t + \varepsilon_t$$
$$\varepsilon_t = -\phi \varepsilon_{t-1} + error.$$

In this case the error term is time-dependent. The symbol *f* represents an autoregressive parameter showing the final estimate of the autocorrelation at lag one.

24.3 Example. The following example may help clarify the AUTOREG procedure. In this example, the dependent variable (VSBY1) is a fog/no fog decision. The predictor variable is the dew point depression (DEWPTDE1).

Table 24-1. SAS AUTOREG procedure.



```
                                         SAS

                                  AUTOREG PROCEDURE


DEPENDENT VARIABLE = VSBY1

                          ORDINARY LEAST SQUARES ESTIMATES

        SSE            1376.248      DFE                 24628
        MSE            0.055881      ROOT MSE            0.236
        SBC            -1130.66      AIC               -1146.88
        Reg RSQ          0.1020      Total Rsq          0.1020
        D-W              0.3069

Variable DF     B Value           Std Error   t Ratio        Prob

INTERCPT 1      0.69719             0.00471   147.948      0.0001
DEWPTDE1 1      0.10178             0.00192    52.884      0.0001

                            Estimates of Autocorrelations from 0.0 ± 1.0

Lag     Covariance     Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  0     0.0558769      1.000000       |                   |******************** |
  1     0.0473108      0.846697       |                   |*****************     |

                            Preliminary MSE=  0.01581902

                     Estimates of the Autoregressive Parameters
               Lag     Coefficient           Std Error             t Ratio
                 1     -0.84669665           0.00339053        -249.723831

                        Unconditional Least Squares Estimates

               SSE               356.663      DFE          24627
               MSE               0.01448      Root MSE     0.120
               SBC              -34375.6      AIC         -34400
               Reg RSQ            0.0031      Total Rsq   0.7673
               D-W               2.1628

Variable  DF    B Value        Std Error      t Ratio         Prob

INTERCPT  1     0.89701        0.00728        123.165        0.0001
DEWPTDE1  1     0.01566        0.00178          8.767        0.0001
A(1)      1     -0.8719        0.00313        -278.87        0.0001
```

a. The AUTOREG procedure first fits a simple linear regression to the data. As seen in the output, the regression model is: y = 0.697 + 0.108x. The D-W statistic for this model is 0.3069, indicating a positive serial correlation. **Note:** R^2 for this model is only 0.102.

b. In the second part of the AUTOREG printout, the procedure fits an autoregressive model to the error term. The model built by the AUTOREG procedure is: y = 0.897 + 0.016x - .872 ZLAG1 (predicted y - actual y). The D-W statistic is 2.163, suggesting no serial correlation exists, and the R^2 has been increased to 0.767.

Chapter 25

FITTING A NONLINEAR MODEL

25.1 Introduction. Can wind speed be used to predict the hour at which fog will dissipate? How does cloud cover affect the surface temperature? How does a decrease in dew point temperature affect the formation of clouds? These are questions about the relationships between pairs of variables: wind speed and fog dissipation, cloud cover and surface temperature, and dew point and cloud cover. As one variable increases, the other variable increases or decreases. Regression analysis uses equations to describe how one variable is related to another variable or group of variables (Schlotzhauer and Littel, 1987). If the relationship is linear, the SAS REG procedure can be used to fit a linear regression model to data. However, frequently the relationship between weather parameters is not well represented by a straight line. In such cases, an analyst must fit another type of model to the data (such as the cumulative Weibull distribution model that is used to model ceiling and visibility distributions). This chapter outlines SAS procedures for fitting nonlinear models.

25.2 Discussion.

a. When we say that a model is linear, we really mean that a straight-line relation exists between the model variables. For example, a straight-line relation between two variables can be summarized with the equation:

$$y = b_0 + b_1 x + \varepsilon. \qquad (1)$$

This equation says that the variable, y, is a function of the variable, x. In addition, the straight line is defined by an intercept parameter, b_0, and a slope parameter, b_1. Also, there is some error in the data, e, since the same x value doesn't always give the same y value. You measure a sample and use the sample to estimate a straight line.

b. Similarly, we say a model is nonlinear if the relation between the model variables is best depicted by a curve. For example, the quadratic regression model is the simplest and most frequently used nonlinear regression model. It may be represented by:

$$y = b_0 + b_1 x + b_2 x^2 + \varepsilon. \qquad (2)$$

This is just a special case of the general linear model since a quadratic regression is still linear in the model parameters, b_n. Another example of a nonlinear model is the cumulative Weibull, given below:

$$y = 1 - e^{(-ax^b)}, \qquad (3)$$

where y and x are data variables and a and b are the parameters to be estimated. The cumulative Weibull model is not a special case of the general linear regression model, but is known as an intrinsically linear model (Draper and Smith, 1981). An intrinsically linear model is a nonlinear model that can be made linear by a transformation of the model parameters.

c. The SAS REG procedure, which uses a least squares regression approach, can be used to estimate the model parameters of linear models. However, when the model parameters are nonlinear, the parameter estimates generally can no longer be obtained so easily. Complicated iterative methods are necessary (Afifi and Clark, 1984). The SAS/STAT User's Guide contains the SAS NLIN procedure which can be used to estimate the model parameters of nonlinear models. The problem with PROC NLIN is that instead of only specifying a list of regressor variables, you also have to specify the derivatives of the model, with respect to the parameters, when using computational iterative methods (Gauss, Marquardt, Newton, Gradient). According to Bevington and Robinson (1992), the Marquardt method is a winner for finding parameter fits most directly and efficiently. The Marquardt method is useful when the parameter estimates are highly correlated. The only derivative-free method used by the NLIN procedure is the secant iterative method.

d. SYSNLIN and MODEL procedures contained in the SAS/ETS software package may be used to estimate nonlinear model parameters. The advantage of the SYSNLIN and MODEL procedures is that the analyst does not have to specify a derivative for the various computational methods used by the procedure (the default Gauss method, and alternate Marquardt method). In normal practice, first run the SYSNLIN and MODEL procedures using the default starting values. This works in many cases. If the parameter estimates do not converge, rerun the procedures using the estimated parameter estimates. Convergence and the rate of convergence may depend on the choice of the starting values for the parameter estimates.

25.3 Examples.

a. Sample SYSNLIN program: The following SAS SYSNLIN code, calculates the parameters (A and B) in a cumulative Weibull distribution equation. The output file is called "TWO" in this example, while Y is the cumulative Weibull distribution function of the visibility (VSBY).

```
DATA ONE;
INPUT Y VSBY;
CARDS;
0.124    0.5
0.243    1.0
0.352    2.0
0.506    3.0
0.642    4.0
0.662    5.0
0.761    6.0
RUN;
PROC SYSNLIN METHOD = MARQUARDT
OUTACTUAL OUTPREDICT OUTRESID
OUT = TWO;
Y = 1 - EXP (-A * VSBY ** B);
PARAMETERS A B;
ENDO Y;
EXO  VSBY;
RUN;
```

b. Sample PROC MODEL program: The SAS code below, for the SAS MODEL procedure, calculates the parameters (A and B) in a cumulative Weibull distribution equation. The parameters and variables are the same as in the previous example. The PROC MODEL code will produce the same output generated by the PROC SYSNLIN code above.

```
DATA ONE;
INPUT Y VSBY;
CARDS;
0.124    0.5
0.243    1.0
0.352    2.0
0.506    3.0
0.642    4.0
0.662    5.0
0.761    6.0
RUN;
PROC MODEL METHOD = MARQUARDT;
PARMS A B;
ENDO Y;
EXO  VSBY;
Y = 1 - EXP(-A * VSBY ** B);
FIT Y/OUTACTUAL OUTPREDICT OUTRESID
OUT=TWO;
RUN;
```

c. Below is an example of the SAS output generated by the previous procedures.

```
                              SYSNLIN Procedure
                                OLS Estimation

                     Nonlinear OLS Summary of Residual Errors


             DF         DF
Equation     Model      Error     SSE       MSE       Root MSE    R-Square    Adj R-Sq

ACTUAL        2          5      0.00339  .0006772    0.02602      0.9900      0.9880

                        Nonlinear OLS Parameter Estimates
              Approx.                          'T'          Approx.
Parameter     Estimate        Std Err         Ratio         Prob>|T|
    A         0.249978        0.02000         12.50         0.0001
    B         0.955560        0.05926         16.13         0.001

      OBS    _ESTYPE_         _TYPE_       _WEIGHT_          Y          VSBY

       1       OLS            ACTUAL          1           0.12400       0.5
       2       OLS            PREDICT         1           0.12094       0.5
       3       OLS            RESIDUAL        1           0.00306       0.5
       4       OLS            ACTUAL          1           0.24300       1.0
       5       OLS            PREDICT         1           0.22118       1.0
       6       OLS            RESIDUAL        1           0.02182       1.0
       7       OLS            ACTUAL          1           0.35200       2.0
       8       OLS            PREDICT         1           0.38417       2.0
       9       OLS            RESIDUAL        1          -0.03217       2.0
      10       OLS            ACTUAL          1           0.50600       3.0
      11       OLS            PREDICT         1           0.51042       3.0
      12       OLS            RESIDUAL        1           0.00442       3.0
      13       OLS            ACTUAL          1           0.64200       4.0
      14       OLS            PREDICT         1           0.60944       4.0
      15       OLS            RESIDUAL        1           0.03256       4.0
      16       OLS            ACTUAL          1           0.66200       5.0
      17       OLS            PREDICT         1           0.68765       5.0
      18       OLS            RESIDUAL        1          -0.02565       5.0
      19       OLS            ACTUAL          1           0.76100       6.0
      20       OLS            PREDICT         1           0.74969       6.0
      21       OLS            RESIDUAL        1           0.01131       6.0
```

Chapter 26

DISCRIMINANT ANALYSIS

26.1 Introduction. Discriminant analysis is a statistical technique that classifies individual observations into groups, and can be used to provide estimates of event probabilities. The idea was conceived in 1936 by R. A. Fisher, but it was Miller (1962) who demonstrated that discriminant analysis could be useful in weather forecasting. These techniques are computationally intensive. High speed computers and the availability of high-level programming languages such as SAS make discriminant analysis a powerful and simple technique for analyzing data and developing forecast models. This chapter outlines some of the basic principles of discriminant analysis, and provides an example of its use.

26.2 Discussion.

a. To illustrate the basic principles of discriminate analysis, consider a simple forecast model consisting of two predictors (X_1 and X_2). On a conventional Cartesian plot (shown above), a closed circle will be plotted for the observed value at each of the predictors when fog is not subsequently observed at the verification hour; an open circle will be plotted when fog is observed. After all the data is plotted, a line that best separates the two categories is drawn. This line, referred to as the "discriminant," serves as the basic forecast model. Subsequent observations of X_1 and X_2 would then be plotted to obtain a forecast. A forecast of "fog" or "no fog" is based upon which side of the discriminant the point lies.

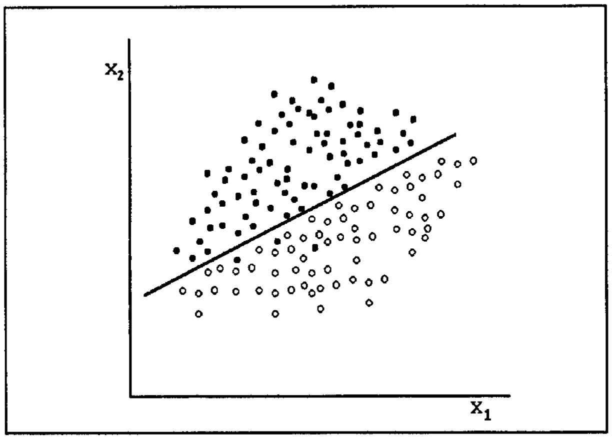


Figure 26-1. Cartesian plot of a forecast model.

b. The SAS/STAT user's guide contains a number of different types of discriminant analysis, which can basically be divided into parametric and a nonparametric forms. There are a number of different subgroups within each form. Parametric discriminant analysis is connected with assumptions about the statistical distribution of the data, while the nonparametric form can be used without making any assumptions about the distribution of the data. The SAS DISCRIM procedure has the capability of doing Miller's strategy, but it is certainly more complicated and experimental than parametric forms. The basic parametric form, linear discriminant analysis, is the only form in which SAS will provide information that can be used to generate an equation to calculate probabilities.

c. Several assumptions are required for linear discriminant analysis, and these assumptions are usually violated. Klecka (1980) made an interesting statement about the violation of these assumptions in discriminant analysis. If the model performs well, the violation of the assumptions is not harmful. Usually attempts to transform the data or use alternative forms provide only marginal improvements.

d. Linear discriminant analysis should be attempted first, before trying more complex forms (such as quadratic discriminant analysis or one of the nonparametric forms). Any number of variables can be used. One variable is defined which indicates the prediction class (i.e., the occurrence or non-occurrence of the variable of interest). For a fog forecast model, the occurrence of fog could be indicated by setting a variable called FOG equal to 1. When no fog occurs this variable has a value of 0 (but we are not restricted to only 2 classes). The model will provide a discriminant function, which is the linear combinations of the predictor variables which will best discriminate between the defined classes. An example of a forecast model using linear discriminant analysis is described in Coffin and Warren (1991).

e. The SAS procedure DISCRIM generates a classification matrix containing numbers that reveal the predictive ability of the discriminant function. The sum of the numbers along the left diagonal (running from the upper left to the lower right) represent the number of correct observations, while the sum of the numbers along the opposite diagonal are the incorrect classifications.

f. To properly evaluate the value of the discriminant function as a forecasting model, a classification matrix must be computed using independent data. This data must not have been used in developing the discriminant function. Usually it is best to simply withhold 1 or 2 years' worth of data to use as an independent data set. The Heidke Skill Score can be used to evaluate the skill of the forecast model.

26.3 Example.

a. Below is a simple example showing how to perform discriminant analysis using SAS. The example uses dew point depression and wind direction to predict the occurrence of fog 3 hours from the observation time.

```
DATA ONE;
INFILE DATAIN MISSOVER;
INPUT YY MM DD HH DDEP WDIR VSBY;
PROC DISCRIM DATA = ONE POOL = YES LIST OUT = TWO;
CLASS VSBY;
VAR DDEP WDIR;
RUN;
```

where:

YY = Year
MM = Month
DD = Day
HH = Hour
DDEP = Dew point depression
WDIR = Wind Direction
VSBY = Visibility (1 = fog / 0 = no fog)

Note: The option POOL = YES is necessary in order to obtain linear discriminant functions. The option LIST gives event probabilities for each individual observation. The option OUT=TWO stores the calibration information in a SAS data set named TWO, which you will need to test the discriminant function on an independent data set. The required CLASS statement defines the group variable.

b. The output from the previous SAS statements (shown below) contains fog probabilities for all the individual observations, a classification matrix, and a linearized discriminant function for each group (fog / no fog).

Table 26-1. Classification matrix

OBSERVED	FORECAST		
	FOG	NO FOG	TOTAL
MISSING	1,123	312	1,435
FOG	5,351 89.57 %	623 10.43 %	5,974
NO FOG	8,962 23.39 %	29361 76.61 %	38,323
TOTAL	15,436	30,296	45,732

Table 26-2. Linearized discriminant function

	Fog	No Fog
CONSTANT	-0.46349	-2.82504
DDEP	0.04727	0.25568
WDIR	0.00711	0.01311

c. Next we provide the SAS code for taking the calibration results of the previous DISCRIM procedure and testing the model on an independent set of data:

```
PROC DISCRIM DATA =TWO TESTDATA
= TEST TESTLIST;
CLASS VSBY;
TESTCLASS VSBY;
VAR WDIR DDEP;
RUN;
```

DATA = TWO represents the calibration information from the previous strategy. The TESTDATA option represents the SAS DATASET containing the independent data (in this case it is named TEST). The TESTLIST option generates group probabilities for all the independent data. The CLASS and TESTCLASS statements represent the grouping variables.

d. In the output, two sets of coefficients are provided: one for FOG and one for NO FOG (in general we can call them Class A and Class B). The equations are then set up in the following manner. The classification matrices obtained from historical and independent data will enable the analyst to learn how the forecasting model is performing. In order to implement the model on future data, the analyst will need to set up the equations that provide probability estimates.

$$A = C_{A0} + \sum_{i=i}^{N} C_{Ai} X_i$$

$$B = C_{B0} + \sum_{i=1}^{N} C_{Bi} X_i$$

$$P(A) = \frac{1}{\exp(B-A) + 1}$$

where C_{A0} and C_{B0} are the constants for class A and B, respectively; the coefficients for the *i* th predictor variable is C_{Ai} and C_{Bi}; and the ith predictor variable has a value of X_i. P(A) represents the probability of class A occurring. N is the total number of variables. Using the example given above, the probability of fog is computed as shown on the next page.

$$\text{FOG} = (0.00711 \bullet \text{WDIR}) + (0.04727 \bullet \text{DDEP}) - 0.046349$$

$$\text{NO FOG} = (0.01311 \bullet \text{WDIR}) + (0.25568 \bullet \text{DDEP}) - 2.82504$$

$$\text{PROB OF FOG} = \frac{1}{\exp(\text{NO FOG} - \text{FOG}) + 1}$$

Thus for a wind direction of 140 degrees and a dew point depression of 1 degree:

$$\text{FOG} = 0.99632$$

$$\text{NO FOG} = -0.73396$$

$$\text{PROB OF FOG} = 0.85$$

In this case, the probability of fog is 85 percent. Evaluation of the skill of the model must be done with an independent dataset. For further details on PROC DISCRIM refer to the SAS/STAT User's Guide, Page 761.

e. The analyst may be interested in knowing how the model is discriminating between fog and no fog. In the case of linear discriminant analysis, this information can be obtained by finding the equation of the linear surface for which the probability of both classes is exactly 50 percent. (For two predictors this surface is a line; for three predictors, a plane; for more than three predictors it is a multidimensional surface referred to as a hyperplane). This equation is known as the Fisher discriminant function. For the two-dimensional case it is given by (Afifi and Clark, 1984):

$$f_1 X_1 + f_2 X_2 = k$$

where f_1 and f_2 are coefficients, X_1 and X_2 are the predictors, and k is a constant. These terms are obtained using the following equations:

$$f_1 = C_{A1} - C_{B1}$$

$$f_2 = C_{A2} - C_{B2}$$

$$k = C_{B0} - C_{A0}\,.$$

This is simply a subtraction of the coefficients in the PROC DISCRIM output. Note that the subtraction to compute k is in the reverse order of the other two subtractions. For our fog forecast model, the Fisher discriminant function is given by:

$$(-0.20841 \bullet \text{DDEP}) - (0.0060 \bullet \text{WDIR}) = -2.36155.$$

The equation of this line is plotted in Figure 26-2.

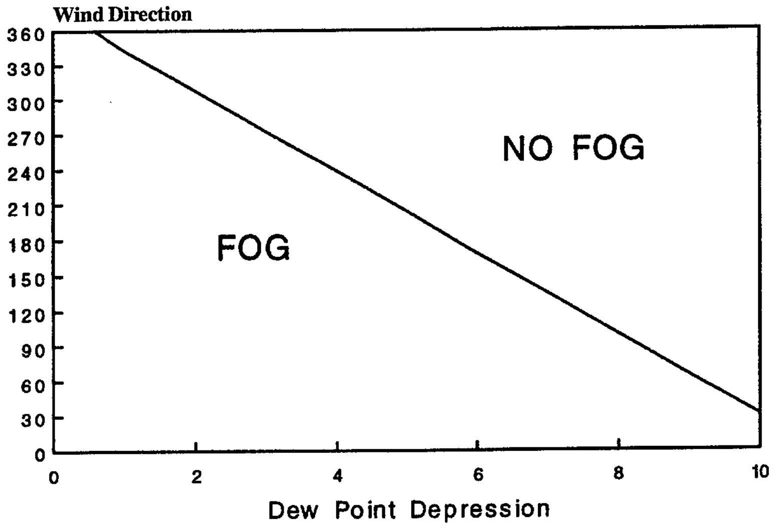


Figure 26-2. Fog forecast model plot.

Note: This graph also illustrates a difficulty in using wind direction as a predictor variable. It is usually better to use the sine or cosine of the wind direction.

Chapter 27

STEPWISE DISCRIMINANT ANALYSIS

27.1 Introduction. Discriminant analysis is a multivariate statistical technique that can be used to develop a mathematical model to estimate event probabilities. Many times it is not known which or how many predictor variables to use in discriminant analysis (SAS DISCRIM procedure). Consequently, analysts collect data on variables that are merely suspected as good discriminators. Stepwise discriminant analysis (SAS STEPDISC procedure) cannot generate a mathematical model, but it is a tool that will enable the analyst to eliminate redundant and unnecessary predictors for the SAS DISCRIM procedure. This will ensure a more efficient model, using the optimum number of predictors.

27.2 Background.

a. Stepwise variable selection procedure is used to select the most useful discriminating variables for group separation in the SAS DISCRIM procedure. This permits the analyst to identify independent predictor variables which contribute the most to the separation of the data. Variables that contribute little, or are redundant, can then be excluded. There are three types of stepwise discriminant analysis used by the SAS STEPDISC procedure: forward selection, backward selection, and stepwise selection.

(1) The forward selection method begins with no variables in the model. With each step, the most discriminatory variables are added one by one. Once a variable has been selected, it stays in the model permanently.

(2) Backward selection is essentially the opposite of forward selection. Initially, all variables are considered and with each step the most unnecessary or least discriminatory variable is eliminated one-by-one. Once a variable has been removed, it is deleted permanently.

(3) Stepwise selection is similar to forward selection, except with each step a variable may be added or deleted. The iteration continues until the procedure identifies the optimum collection of variables. This method is the SAS STEPDISC procedure default.

b. Forward selection is generally the easiest of the three to understand. Its use for meteorological applications is reported in the literature. Forward selection ensures that variables entered are not subsequently removed; this may be desirable in some applications. However, forward selection does not ensure that the first two variables selected are necessarily the best pair.

c. Thompson and Zucchini (1990) point out that simply increasing the number of predictor variables in a model does not necessarily improve the accuracy of the model forecasts. In fact, excessive complexity usually decreases the value of a particular model. Miller (1962) used forward discriminant analysis in early forecast model to analyze 175 possible predictor variables. His technique identified that only 16 of these variables were necessary to predict precipitation conditions at Hartford, Conn., over an interval from zero to 6 hours in advance.

27.3 The SAS STEPDISC Procedure.

a. Evaluation of potential predictors. To determine which variables should be added or deleted from the set of predictors, SAS uses statistical measures such as the squared partial correlation, Wilks' lambda, and a tolerance test. The first two measures are converted to an F-statistic and a probability level. Variables are entered or deleted depending on the size of the F-value and respective probability levels.

b. Forward, Backward, or Stepwise? Afifi and Clark (1984) state that unless the analyst is familiar with the complexities of a given option, the stepwise option should be selected. Another important decision the analyst must consider is what value to use for the acceptance/rejection threshold. Afifi and Clark recommend as a threshold the particular F-value that corresponds to a probability value of 0.15. This is the SAS default. However, other researchers have used different values.

c. The output from the STEPDISC procedure provides the information needed to determine which variables should be included in the DISCRIM procedure. The easiest and most direct way to evaluate the variables is to look at the average squared canonical correlation (ASCC). The ASCC, which ranges from zero to one, indicates the percentage of the variance which is explained by the predictor variables. An ASCC of 1 indicates that all groups (e.g. fog and no fog) are perfectly separated with the variables considered. An ASCC of 0 indicates that the groups are not separated, i.e., it is not possible to discriminate between the groups using the input variables. The output of the STEPDISC procedure will list the variables in order from the most discriminatory to the least. At some point, the ASCC values stabilize, which is the cue that additional variables are adding very little, if any, additional skill to the model. All variables up to this point should then be selected for the DISCRIM procedure.

27.4 Example.

a. Shown below is an example of how to perform stepwise discriminant analysis using SAS. The example develops a discriminant analysis model to forecast fog with a visibility of less than or equal to 5,000 meters (yes/no) 3 hours from the observation time. Potential predictors from the current observation include: temperature (TEMP), dew point (DEW), dew point depression (DDEP), wind direction (WDIR), cosine and sine of the wind direction (CWD and SWD), wind speed (WSP), cosine and sine of the hour angle (CHH and SHH), altimeter setting (ALT), sea level pressure (SLP), ceiling height (CIG), and visibility (VSBY). The class variable VSBY3 is set equal to 0 when fog (visibility less than or equal to 5000 m) occurs 3 hours from the observation point, otherwise it has a value of 1. (The variables YY, MM, DD, HH are the year, month, day, and hour of the observation; we did not include these as potential predictors.) The SAS code is as follows:

```
DATA ONE;
INFILE DATAIN MISSOVER;
INPUT YY MM DD HH CHH SHH CWD SWD TEMP DDEP WSP DEW WDIR
CIG SLP ALT VSBY VSBY3;
PROC STEPDISC;
CLASS VSBY3;
VAR CHH SHH CWD SWD TEMP DDEP WSP DEW WDIR CIG SLP ALT VSBY;
RUN;
```

b. A summary of the output from this procedure is shown below.

Stepwise Discriminant Analysis

	Variable				Average
	DEW				Squared
Step	Entered	Partial	F	Prob >0	Canonical
		R**2	Statistic	F	Correlation
1	VSBY	0.4242	31886.274	0.0001	0.42417035
2	DDEP	0.0420	1896.459	0.0001	0.44833985
3	SLP	0.0222	988.032	0.0001	0.46059021
4	SWD	0.0124	544.656	0.0001	0.46729342
5	TEMP	0.0095	414.288	0.0001	0.47234394
6	CIG	0.0036	154.549	0.0001	0.47422135
7	WDIR	0.0024	105.395	0.0001	0.47549859
8	CHH	0.0011	47.885	0.0001	0.47607826
9	WSP	0.0006	26.393	0.0001	0.47639757
10	CWD	0.0005	19.808	0.0001	0.47663711
11	ALT	0.0004	16.417	0.0001	0.47683557
12	SHH	0.0001	3.787	0.0517	0.47688135

c. In the upper part of the table, the variables that failed the tolerance test are depicted. In this example only one variable, the dew point, failed this test. The other 12 variables are listed in the lower part of the table. The most discriminatory variable is the visibility. The average squared canonical correlation (ASCC) of this variable alone is 0.4242. The addition of the variables dew point depression, sea level pressure, the sine of the wind direction, and temperature, increase the ASCC to 0.4723. At this step, the ASCC value stabilizes. The addition of the fifth variable (TEMP) increases the ASCC by about 0.5 percent. The addition of the sixth variable (CIG) increases the ASCC by only 0.2 percent. In fact, the ASCC with all twelve variables is only 0.4769, an increase of only 0.5 percent from just using five variables. Thus, this analysis suggests that using only the first five predictors will discriminate between fog and no fog.

27.5 Conclusion. Stepwise discriminant analysis is a technique for identifying which variables are best used as predictors in developing a forecast model. It can eliminate redundant and unnecessary variables. The average squared canonical correlation is a useful tool for analyzing the marginal improvement to a model each additional variable provides.

Chapter 28

A COMPARISON OF DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

28.1 Introduction. This chapter describes some of the similarities and differences between logistic regression and discriminant analysis (described in a previous paper). Both of these techniques are useful for analyzing binary response data (yes/no occurrences). The idea for this paper came from an article presented at the *12th Conference on Probability and Statistics in Atmospheric Sciences* (1992), in which logistic regression and discriminant analysis techniques are compared for forecasting thunderstorm occurrences.

28.2 Discussion.

a. Discriminant Analysis. In linear discriminant analysis, the SAS DISCRIM procedure generates linear discriminant functions for each class of the response variable (such as “thunderstorms” vs. “no thunderstorms”).

$$\text{TSTM} = b_1\text{Var}_1 + b_2\text{Var}_2 + \ldots + b_n\text{Var}_n - b_o$$

$$\text{No TSTM} = b_1\text{Var}_1 + b_2\text{Var}_2 + \ldots + b_n\text{Var}_n - b_o$$

where the coefficients of the predictor variables are denoted by b_1, b_2, b_n, and b_o denotes a constant. The probability of No TSTMS is

$$\frac{1}{\exp(\mathit{No\ TSTM} - \mathit{TSTM}) + 1}.$$

b. Logistic regression. As mentioned previously, logistic regression is an alternative to discriminant analysis. Logistic regression provides a means of fitting an S-shaped curve to data in which the dependent variable is binary (TSTM / No TSTM).

For the logistic regression model the probability of No TSTMS can be expressed as follows:

$$\frac{1}{1 + \exp[-(b_o + b_1\text{Var}_1 + b_2\text{Var}_2 + \ldots + b_n\text{Var}_n)]}.$$

In both cases the probability of TSTM equals one minus the probability of No TSTM.

c. Differences between discriminant analysis and Logistic regression. During a workshop referenced at the statistics conference, there was found to be little difference between the results obtained when analyzing the TSTM data with discriminant analysis and logistic regression. Even though there may be little practical difference between the two techniques, there are some differences in the underlying assumptions of each. One of the primary assumptions of discriminant analysis is normality. If the data being analyzed is assumed to be normal, then it is preferable to use discriminant analysis instead of logistic regression. In most applications of discriminant analysis at least one variable is qualitative or dichotomous (0 or 1). The assumption of multivariate normality will rarely be satisfied if dichotomous predictor variables are present. Application of the discriminant function when the assumption of normality does not hold may result in bias. Logistic regression, on the other hand, is applicable for any combination of discrete or continuous variables.

28.3 Sample Calculations. The following examples demonstrate how to use discriminant analysis and logistic regression within SAS. Some definitions of the SAS options that highlight the differences between the two techniques are also provided.

CHAPTER 28

a. The SAS code used to run discriminant analysis and logistic regression for the "Eglin Thunderstorm" project (Cornell, 1993) is shown below:

```
*       ENTER TSTM DATA
*
*       VARIABLES:
*               CCL - CONVECTIVE CONDENSATION LEVEL
*               PLCL - PRESSURE AT THE LCL
*               PLFC - PRESSURE AT THE LFC
*               DPD - DEWPOINT DEPRESSION
*               SW - SHOWALTER INDEX
*               TIME - 1= NO TSTM,  0=TSTM
*;

DATA ONE;
INPUT CCL PLCL PLFC DPD SW TIME;
CARDS;
49      990     986     26      10      1
40      957     618     25      2       1
151     862     100     17      14      1
...(ETC)
RUN;

*
*       DISCRIMINANT ANALYSIS
*;



DATA TWO;
SET ONE;
PROC DISCRIM POOL=YES LIST OUT=PRED;
CLASS TIME;
VAR CCL PLCL PLFC DPD SW;
RUN;
*
*       LOGISTIC REGRESSION
*;
DATA THREE;
SET ONE;
PROC LOGISTIC;
MODEL TIME = CCL PLCL PLFC DPD SW / CTABLE;
OUTPUT OUT = PRED P=PHAT LOWER=LCL UPPER=UCL;
RUN;
```

b. Output from the logistic regression procedure:

LOGISTIC EQUATION	PARAMETER ESTIMATE
Intercept	-7.930
CCL	-0.007
PLCL	0.009
PLFC	0.001
DPD	-0.020
SW	-0.207

CRITERIA FOR ASSESSING MODEL FIT
SCORE -178.697 (p=0.0001)

PROBABILITY TABLE

OBS	CCL	PLCL	PLFC	DPD	SW	TIME	LEVEL	PHAT	LCL	UCL
1	49	990	986	26	10	1	0	0.401	0.253	0.57
2	40	957	618	25	2	1	0	0.654	0.593	0.711

CLASSIFICATION TABLE

	CORRECT		INCORRECT		% CORRECT	SENSITIVITY	SPECIFICITY	FALSE POS.	FALSE NEG.
PROB LEVEL	EVENT	NON EVENT	EVENT	NON EVENT					
0.48	389	145	127	45	75.6	89.6	53.3	24.6	23.7
0.5	383	148	124	51	72.2	88.2	54.4	24.5	25.6

c. Output from the discriminant analysis procedure:

LINEAR DISCRIMINANT FUNCTIONS

Variable	THUNDERSTORM (0)	NO THUNDERSTORM (1)
Constant	-757.584	-749.202
CCL	0.999	1.000
PLCL	1.493	1.483
PLFC	0.022	0.020
DPD	-0.071	-0.049
SW	0.344	0.536

PROBABILITY OF SAS DISCRIM

FROM TIME	CLASSIFIED INTO TIME		0	1
1	1		0.3996	0.6004
1	0	*	0.5902	0.4098
1	1		0.0283	0.8717

where * denotes misclassified observation
0 denotes thunderstorm
1 denotes no thunderstorm

CLASSIFICATION SUMMARY

FORECAST

		TSTM	NO TSTM	TOTAL
OBSERVED	TSTM	354	80	434
	NO TSTM	107	165	272
	TOTAL	461	245	706

d. Explanation of SAS code options.

(1). In the SAS DISCRIM procedure, the option POOL=YES is used to compute the linear discriminant functions. The LIST option prints the classification results for each observation. The CLASS statement, which is required in the DISCRIM procedure, defines the groups, TSTM or NO TSTM, in the present example.

(2). In the SAS LOGISTIC procedure, the CTABLE option prints the classification table for the final model (the CTABLE option is only available for binary response data). The PPROB option specifies the critical probability value in classifying observations for the CTABLE option. The PPROB must be between 0 and 1. For the classification table, the response is predicted to be an event (yes) if the estimated probability value is greater than zero, or equal to the value stated by the PPROB option. Otherwise, the response is predicted to be a nonevent (no). PPROB is set to .05 by default. The PPROB option is ignored if the CTABLE option is not specified.

e. Explanation of the SAS code output.

(1) DISCRIM Procedure.

The Fisher Discriminant Function (Z) parameters are obtained by subtracting the values of the coefficients and constants of the linear discriminant functions for the TSTM case from those for the NO TSTM case. In our example this is

$$Z = 0.192(SW) + 0.022(DPD) - 0.002(PLFC) - 0.010(PLCL) + 0.001(CCL) + 8.382\,.$$

The probability of NO TSTM is

$$\frac{1}{1+\exp(\mathrm{Z})}$$

and the probability of TSTM is one minus the probability of NO TSTM.

(2) LOGISTIC Procedure.

For logistic regression, the probability of NO TSTM is

$$\frac{1}{1+\exp\{-[-7.930-0.007(CCL)+0.009(PLCL)+0.001(PLFC)-0.020(OPD)-0.207(SW)]\}}\,.$$

Again, the probability of TSTM is one minus the probability of NO TSTM.

For logistic regression, the SCORE statistic under "Criteria For Assessing Model Fit" provides a test of the joint significance of the predictor variables. Since the probability value is very low ($p < 0.05$), the predictors are deemed to be useful in the model.

In the logistic regression output, "Time" can either be 0 or 1. For observation 1, PHAT=0.401 means that the probability that Time=0 (TSTM) is 0.401. UCL and LCL are the upper and lower confidence limits for the probability PHAT. The _LEVEL_ (a SAS variable) will always be zero if the Time variable has two possibilities (0 or 1).

28.4 Evaluation of Results. The following measures provide a method to compare the results of the two techniques.

a. Discriminant Analysis.

FORECAST

OBSERVED

	YES	NO	SUM
YES	354	80	434
NO	107	165	272
SUM	461	245	706

FY = Forecast Yes; FN = Forecast No; OY = Observed Yes; ON = Observed No

Percent Correct = (FYOY + FNON) / (Sum OY + Sum ON) = 73.51 percent

False Positive = (FYON) / (SUM FY) = 0.2321

False Negative = (FNOY) / (Sum FN) = 0.3265

Sensitivity = (FYOY) / (Sum OY) = 0.8157

Specificity = (FNON) / (Sum ON) = 0.6066

Critical Success Index = (FYOY) / (FYOY + FNOY + FYON) = 0.6543

True Skill Score = (FYOY / Sum OY) - (FYON / Sum ON) = 0.426

b. Logistic Regression.

FORECAST

OBSERVED

	YES	NO	SUM
YES	383	51	434
NO	124	148	272
SUM	507	199	706

FY = Forecast Yes; FN = Forecast No; OY = Observed Yes; ON = Observed No

Percent Correct = (FYOY + FNON) / (Sum OY + Sum ON) = 75.21 percent

False Positive = (FYON) / (SUM FY) = 0.2446

False Negative = (FNOY) / (Sum FN) = 0.2563

Sensitivity = (FYOY) / (Sum OY) = 0.8825

Specificity = (FNON) / (Sum ON) = 0.5441

Critical Success Index = (FYOY) / (FYOY + FNOY + FYON) = 0.6864

True Skill Score = (FYOY / Sum OY) - (FYON / Sum ON) = 0.4264

c. The false positive rate is the proportion of predicted "yes" events that were observed to be "no" events. The false negative rate is the proportion of predicted "no" events that were observed as "yes" events. The sensitivity is the proportion of observed "yes" events that were predicted "yes" events. The specificity is the proportion of observed "no" events that were predicted "no" events. The critical success index (CSI) takes into account both classes of error, failures to predict and false alarms. The CSI varies from 0 (total failure) to 1 (perfection). The true skill score (TSS) provides a measure of the ratio of the observed skill to perfect skill. TSS ranges from -1 to +1.

d. Not all of the measures shown above are provided automatically by the SAS procedures, some were calculated by hand. The SAS DISCRIM procedure provides a classification table and error count estimates. The SAS LOGISTIC procedure provides percent correct, false positive, false negative, sensitivity, and specificity.

28.5 Conclusion. There is not much difference between the results of the two techniques in this example, but there are a few differences in the output formats. The SAS LOGISTIC procedure provides more evaluation measures than the DISCRIM procedure, and it is slightly easier to formulate the probability equation with the LOGISTIC procedure. The probability table in the LOGISTIC procedure is more complicated than the one from DISCRIM, but it provides upper and lower confidence limits for the estimated probability.

Chapter 29

THE CENTRAL LIMIT THEOREM

29.1 Introduction. The Central Limit Theorem is one of the more influential theorems in statistics. This chapter provides a definition of the Central Limit Theorem, and describes some viewpoints that have been collected from a literature survey.

29.2 Discussion.

a. Definition of the Central Limit Theorem. The mean of a population is generally estimated from a sample of observations. As the random sample size (n) is increased (i.e., $n \rightarrow \infty$), the distribution of the sample means approaches a normal distribution regardless of the shape of the parent population. In the literature, the approximation is generally accepted as sufficient when the sample size is at least 30. Even if the data distribution is far from normal, the distribution of sample means tends toward a normal distribution as the sample size increases. This fact is probably the single most important reason for the widespread use of the normal distribution.

b. Example. Table 29-1 shows mean January cloud-cover percentages.

Table 29-1. Mean January cloud-cover percentages.

YEAR	%	YEAR	%	YEAR	%	YEAR	%
1960	0.22	1968	0.12	1976	0.12	1984	0.11
1961	0.18	1969	0.28	1977	0.36	1985	0.18
1962	0.27	1970	0.06	1978	0.18	1986	0.12
1963	0.31	1971	0.26	1979	0.23	1987	0.08
1964	0.07	1972	0.33	1980	0.22	1988	0.18
1965	0.11	1973	0.15	1981	0.06	1989	0.16
1966	0.14	1974	0.2	1982	0.18	1990	0.28
1967	0.28	1975	0.17	1983	0.29	1991	0.43
Mean = 0.20				Standard Deviation = 0.09			

Observations of cloud-cover generally follow a very non-normal distribution. However, since the sample size is greater than 30, we can invoke the Central Limit Theorem and assume that the distribution of mean cloud cover is normal. A confidence interval on the mean can be readily calculated.

c. Bradley (1973) states that some of the earlier investigators of the Central Limit Theorem concluded that the distribution of the sample mean becomes normal even for very small sample values, regardless of the shape of the distribution. However, this is true only if the sample population is not appreciably skewed. The more skewed the population, the greater the sample size must be to invoke the Central Limit Theorem.

d. Bradley (1973) warns that the Central Limit Theorem may be inappropriately invoked in some cases where the population data is greatly skewed. In extreme cases, it can be necessary to have sample sizes on the order of hundreds or even thousands before the Central Limit Theorem becomes valid. Kurtosis in the population distribution can also invalidate the Central Limit Theorem for small sample sizes. (Kurtosis is another measure of the shape of the distribution of values. Large values of kurtosis indicate the distribution has "heavy tails.")

e. The Central Limit Theorem can be used to answer the question, "Is the sample mean a good estimate of the population mean?" Suppose one computes the means of a group of samples, and constructs a frequency distribution of these means. The dispersion of the distribution of means is given by

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where $s_{\bar{x}}$ is the standard deviation of the distribution of sample means (also called the standard error), and s is the standard deviation of the original population from which a sample of size n is drawn. For the example above, the 95 percent confidence interval (at the 5 percent level of significance) is

$$0.20 \pm 2.04 \left(\frac{0.09}{\sqrt{32}} \right) = 0.20 \pm 0.03.$$

Chapter 30

STATISTICAL INTERVALS

30.1 Introduction.

a. AFCCC (AWS), often provides customers with single values (such as mean temperatures, or ASPAM vertical profile parameters) to answer a request. However, some customers might be better served if AFCCC provided a degree of confidence in that single value, and a range of practical values. For example, suppose the threshold for requiring air conditioning at a base is a mean monthly temperature greater than 80° F. The base engineers are notified by weather that the mean monthly temperature for a particular month is 78° F. Based on the 80° F threshold, the base engineers would not turn on the air conditioning for that month. However, if weather instead tells the customer that with 95 percent confidence the mean monthly temperature is 78 ± 2.5° F, the base engineers might make the decision to turn on the air conditioning.

b. This chapter presents a general definition of a confidence interval, and gives some examples of the applications of confidence intervals. This chapter also describes how to calculate prediction intervals, tolerance intervals, error bars, and to predict future observations. One possible application of special importance is the construction of confidence intervals for vertical profiles produced by the Atmospheric Slant Path Analysis Model (ASPAM).

30.2 Background.

a. A population is the totality of elements under study (i.e., all observations), whereas a sample includes only a portion of the population. Parameters describe the properties of populations, while statistics describe samples. Generally, we use random sample statistics to infer characteristics about a population. Examples of parameters include the population mean, μ, and the population standard deviation, σ. Statistics include the sample mean, $\bar{x}$, and the sample standard deviation, *s*.

b. A confidence interval (abbreviated C.I.) for a population parameter gives an interval estimate for the parameter. The estimate places upper and lower bounds (i.e., confidence limits) around a point estimate for the parameter. Sample size (n) and population variability (s) affect the precision of the estimate (Schlotzhauer and Littel, 1987).

c. The confidence level is the probability that an assertion about the value of a population parameter is correct. That is, it indicates your degree of belief that the interval contains the true population parameter.

d. A normal distribution is one of many theoretical distributions for a population. Many statistical methods assume the values in a data set are a sample from a normal distribution. The normal distribution is completely defined by its mean μ and standard deviation σ. However, in most real-life applications, σ is unknown so the sample standard deviation s is used. Once you replace s with the sample standard deviation *s*, using the normal distribution is not exactly correct. A t-distribution should be used instead. A t-distribution is very similar to the normal distribution and allows you to adjust for different sample sizes. The t-value is based both on the sample size and on the level of confidence you choose. If the sample values are not normally distributed, the t-values can be adjusted to allow for departure from normality (Schlotzhauer and Littel, 1987).

30.3 Discussion of Various Statistical Intervals.

a. Confidence Intervals.

(1) The purpose of confidence intervals is to determine a range of values to estimate an unknown population parameter, such as a population mean. For example, if we compute a mean and standard deviation s from a large sample, a confidence interval of the population mean (with a 95 percent level of confidence) is given by:

$$\bar{x} - t_{0.025}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025}\frac{s}{\sqrt{n}} \qquad \textbf{(1)}$$

where μ is the population mean. The value of the confidence interval is based on the sample size and the level of confidence you choose.

Interpretation: There is a 95 percent probability that the population mean falls within this confidence interval. Specifically, if you collect a great many samples and calculate a 95 percent C.I. about μ for each sample, 95 percent of the confidence intervals would contain the true population mean μ, and 5 percent would not. Unfortunately, with only one sample, you don't know whether the C.I. you calculated is one of the 95 percent or one of the 5 percent (Schlotzhauer and Littel, 1987).

(2) Confidence intervals can be applied to:

- Means
- Difference between two means
- Ratio of two means
- Correlation coefficients
- Medians
- Difference between two medians
- Ranges
- Regression coefficients
- Standard deviations
- Rare events
- Individual observations

b. Prediction intervals.

(1) Prediction intervals can also be used to predict a future individual observation from a population. If the sample mean and variance are known in a sample of size n, then the interval for the prediction of a random observation, designated x_{n+1} is

$$\bar{x} - t_{\alpha/2}\sqrt{s^2\left(\frac{n+1}{n}\right)} \le x_{n+1} \le \bar{x} + t_{\alpha/2}\sqrt{s^2\left(\frac{n+1}{n}\right)} \qquad (2)$$

where:

$t_{a/2}$ = t-value at the (1 - a) level of confidence with n -1 degrees of freedom
n = sample size
$\bar{x}$ = sample mean
s^2 = sample variance

(2) The regression equation relating x and y can be used to predict a value of y for a given value of x. Prediction intervals can be used to indicate a likely range of the predicted values of y.

(3) For an individual y value, the interval is called the prediction interval.

(4) The prediction interval for an individual value y_i for a given value x_i is

$$y_i \pm (t_{\alpha/2}\ s_{y\bullet x})\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (3)$$

where:

y_i = predicted value from the regression equation
$t_{a/2}$ = t-value at (1 - a) level of confidence
$\bar{y}$ = sample mean of y-values
$\bar{x}$ = sample mean of x-values
x_i = individual x-values

and

$$s^2_{y\bullet x} = \frac{1}{n-2}\left\{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}. \qquad (4)$$

c. Error bars.

(1) According to Panofsky and Brier (1968), an error bar ($\bar{e}$) is defined by:

$$\bar{e} = \frac{\sum_{i=1}^{n}|F_i - O_i|}{n} \qquad (5)$$

where:

F_i = is the forecast value
O_i = is the observed value
n = the number of observations.

(2) An error bar requires knowledge of the forecasted values.

d. Wilks distribution-free tolerance limits .

(1) Tolerance limits specify the limits within which a certain portion of the population can be expected to occur (with a preassigned probability or level of confidence). In other words, suppose we wish to determine the interval in which a certain percentage, γ, of the population values occurs with confidence 1 - α and provides us with the required random sample size, n. The random sample maximum and minimum values x_{max} and x_{min} are the tolerance limits, and we can conclude (with a confidence level of 1 - α) that g percentage of the population occurs within this range. The Wilks equation for tolerance limits is

$$n\gamma^{n-1} - (n - 1)\gamma^{n} = \alpha . \qquad \textbf{(6)}$$

(2) Suppose we wish to determine an interval in which 90 percent of the observations occur within the population with a confidence level of 95 percent (the confidence level is 1 - α, so α = 0.05). Equation (6) can be solved by trial and error. A value of n = 46 results in

$$(46)(0.9)^{45} - (45)(0.9)^{46} = 0.40 - 0.35 = 0.05. \quad (7)$$

Thus, 90 percent of the observations in a given population lie in the interval determined by the largest and smallest value of a random sample of 46 observations drawn from the population, with a confidence level of 95 percent.

e. ASPAM atmospheric profiles.

(1) In the analysis of meteorological data valid at a point in time, analysts generally have only one observational value for the data point in question. For example, a particular ASPAM-derived profile has only one observation for each meteorological parameter at each atmospheric level. The types of confidence intervals described above apply only when several samples are available. Specifying confidence limits with only one observation value, requires making several assumptions, such as the nature of the variance, standard deviation, and mean of the true population. The method selected also depends upon the question being answered.

(2) The purpose of the first example is to determine a range of values in which the true temperature lies, given the one measurement. In other words, does the Optimum Interpolation Vertical Profile (OIVP) (or RAOB Vertical Profile) temperatures make sense based on the sampling method used.

(a) First, the analyst must obtain estimates of the observational error. This error can result from a variety of sources. Some possible sources of error include statistical (or sampling) error, instrument error, and interpolation error. For ASPAM, since there is one observation in each sample (i.e., one vertical profile), these errors are not calculated from the sample, but are empirically (i.e., experimentally) derived. So, the analyst must first assume and trust these empirical mean, variance, and standard deviation values are the true population parameters. If one also assumes these error sources are independent of each other, the total error, s_T^2, is given by (Bevington and Robinson, 1992):

$$s_T^2 = s_1^2 + s_2^2 + s_3^3 + \ldots \qquad \textbf{(8)}$$

where $s_1^2, s_2^2, s_3^2, \ldots$ are the variances of the individual sources of error.

(b) Assume the temperature at a given point is estimated to be 61° F. Estimates of the standard deviation of the error components are: instrument 1.3° F, interpolation 2.7° F, and statistical 0.5° F. (Recall, the standard deviation is the square root of the variance.). If one assumes that the errors are normally distributed about the true value, then 68 percent of the time, the actual temperature will lie within 1 standard deviation of the observed value. The standard deviation from all sources is then

$$s_T = \sqrt{(1.3)^2 + (2.5)^2 + (0.5)^2} \qquad \textbf{(9)}$$

$$= 2.9 \qquad .$$

(c) If we assume that the observed value of 61° F is the best estimate of the true temperature, then the range of values in which the true temperature lies is

$$61 \pm 2(2.9) = 55.2 \text{ to } 66.8. \qquad \textbf{(10)}$$

In this example, there is a 95 percent probability that the true temperature lies within a range of 55.2° F through 66.8° F. In other words, 5 percent of the time the actual temperature will lie outside of this interval.

(d) The difficulty in using this method lies in attempting to obtain reasonable, trustworthy estimates of the variances of the various components of error. These errors are dependent upon the characteristics of many independent, randomly-drawn samples, and are generally functions of location, altitude, and sensor type. The interpolation error, for example, is dependent upon the number of nearby reporting stations, the terrain, etc. However, since there is only one sample of one observation for each ASPAM vertical profile, one must use some experimentally predetermined errors.

3) As a second example, suppose an analyst wishes to determine if the temperature measurement is consistent with climatology. Suppose, for the same month and hour, there are 18 previous measurements of temperature. These measurements are all independent of each other. The mean value of these samples is 48° F, and the standard deviation is 7° F (T = 48.0° F, s = 7.0° F, n = 18). The current temperature measurement is 61° F and the analyst wants to determine if the value is statistically consistent with the past samples. Our null hypothesis for this test is

H_0: The current observation is drawn from a population with the same mean and standard deviation as the sample mean and standard deviation.

To evaluate this hypothesis we use equation 2. If we assume a 95 percent prediction interval, 17 degrees of freedom (n - 1), and a t-value of 2.11, then equation 2 can be written as equation 11.

$$48 - (2.11)\sqrt{(7.0)^2\frac{19}{18}} \le x_{n+1} \le 48 + (2.11)\sqrt{(7.0)^2\frac{19}{18}}$$

$$48 - 15.2 \le x_{n+1} \le 48 + 15.2 \qquad \textbf{(11)}$$

$$32.8 \le x_{n+1} \le 63.2$$

Since the single random observation of 61° F falls within the 95 percent prediction interval of 63.2 and 32.8, we accept the null hypothesis.

30.4 Conclusion. Statistical intervals have numerous definitions depending on the application. That is, the term "statistical interval" can take on a wide range of meanings depending on context. To avoid confusion in the analysis of data, the user should not arbitrarily select one type of interval. It is better to analyze problems in terms of what specific information is desired, or what hypothesis is to be tested. The appropriate statistical test is then selected to provide this information, or to test the hypothesis.

Chapter 31

CONFIDENCE INTERVALS FOR POPULATION MEANS

31.1 Introduction. A number of SAS procedures in the SAS/STAT user's guide provide confidence intervals for different sample statistics. This chapter focuses on using confidence intervals to analyze estimates of population means.

31.2 Discussion.

a. The mean is often used as a measure of the central value of a sample, while the standard deviation is used to describe the scatter about the mean. The equations for calculating the mean and standard deviation are

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\frac{1}{n - 1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i represents the individual observations, and n is the sample size.

b. The standard error is a standard deviation of a distribution of random sample means. If repeated samples of size n are drawn from a population, the central limit theorem states that the distribution of sample means approaches a normal distribution, even though the population is not necessarily normally distributed. The distribution of the series of sample means has population mean μ and a standard deviation (or standard error of the mean) given by

$$Standard\ Error = \frac{s}{\sqrt{n}}.$$

The standard error measures the spread of a series of random sample means, and can therefore be used to make inferences about the likelihood that the population mean lies within a specified interval. In other words, confidence limits can be calculated by using the standard error. Three formulas for calculating upper- and lower-confidence intervals are shown below. A 95 percent confidence limit is used for all these examples.

(1). For sufficiently large samples (sample size greater than 30) the normal approximation can be used:

$$\left(\bar{x} - 2\frac{s}{\sqrt{n}}\right) \le \mu \le \left(\bar{x} + 2\frac{s}{\sqrt{n}}\right).$$

This indicates the population mean lies within ±2 standard errors of the sample mean.

(2). For small sample sizes (less than 30), the t-distribution can be used as follows:

$$\left(\bar{x} - t_{.05}\frac{s}{\sqrt{n}}\right) \le \mu \le \left(\bar{x} + t_{.05}\frac{s}{\sqrt{n}}\right)$$

where $t_{.05}$ is the value of the (two-sided) t-distribution with n-1 degrees of freedom (this value can be determined from a t table in any statistics book).

(3). If the sample size is small and there are doubts as to whether or not the population is normally distributed, the Chebyshev formula can be used to estimate the 95 percent confidence intervals:

$$\left(\bar{x} - \frac{4.5s}{\sqrt{n}}\right) \le \mu \le \left(\bar{x} + \frac{4.5s}{\sqrt{n}}\right).$$

The Chebyshev formula states that the true population mean will always lie within ±4.5 standard errors of the sample mean (with a 95 percent confidence level).

31.3 Example. Suppose an analyst wishes to estimate the true me an monthly temperature at a particular location. The random sample mean is calculated from the mean monthly temperatures measured for 36 different years (n = 36). If this mean is 53.25, and the standard deviation is 17.70, then the standard error is

$$Standard\ Error = \frac{17.70}{\sqrt{36}} = 2.95.$$

Assuming the population is normally distributed, the analyst can use the normal approximation to estimate the confidence interval of the true population mean monthly temperature as follows:

$$53.25 - 2(2.95) \leq \mu \leq 53.25 + 2(2.95)$$

$$47.4 \leq \mu \leq 59.2.$$

The equation shows, with 95 percent confidence, that the population mean lies between 47.4 and 59.2. If analysts are not sure that the population is normally distributed (or if the actual distribution is unknown), they can use the Chebyshev formula to calculate the confidence interval as follows:

$$53.25 - 4.5(2.95) \leq \mu \leq 53.25 + 4.5(2.95)$$

$$40.0 \leq \mu \leq 66.5.$$

In this case the confidence is wider.

Chapter 32

FORECAST VERIFICATION MEASURES

32.1 Introduction. Many techniques are used to evaluate categorical forecasts. This chapter defines and discusses several of the most commonly used methods.

32.2 Discussion.

a. The Mean Square Error. Accuracy is the primary criterion used for selecting a forecasting model. Accuracy is defined as the degree of correspondence between individual forecasts and observations. Perhaps the most widely used measure of accuracy is the mean square error (MSE), which is calculated by squaring the difference between predicted values and actual observations, then averaging the squared values. Mathematically, MSE is

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(f_i - e_i)^2 \qquad (1)$$

where f_i is the predicted observation, e_i is the actual observation, and n is the number of observations. The root mean square error (RMSE) is the square root of the MSE:

$$RMSE = \sqrt{MSE}. \qquad (2)$$

b. The Brier Score. The Brier score is a measure of accuracy commonly used for probability forecasts. The Brier score is a mean square error method of measuring the accuracy of probability forecasts. The National Weather Service uses the Brier score as a yardstick to measure the accuracy of probability forecasts. The Brier score is 1/2 of the P-score as defined by Brier (1950) and Panofsky and Brier (1958):

$$P = \frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n}(f_{ij} - e_{ij})^2 \qquad (3)$$

where for n occasions, an event can occur in only one of r possible classes. The notation f_{ij} represents the forecast probability. The notation e_{ij} represents the actual occurrence, which can take only the values 0 or 1, according to whether the event occurred or not. For perfect forecasting the P-score will have a value of zero, and for the worst possible forecasting it will have a value of two. The Brier score then has a range of zero to one. Sometimes in the literature the Brier score is referred to as the Half-Brier score, but these terms are identical. The Brier score can be interpreted as the MSE of probability forecasts.

(1). Below is an example of an evaluation of a set of 10 probability forecasts for the occurrence of rain (the data is shown in Table 32-1, below). In the "RAIN" column, $e_{ij} = 1$ when rain occurred, and $e_{ij} = 0$ when rain did not occur. In the "NO RAIN" column, $e_{ij} = 0$ when rain occurred, and $e_{ij} = 1$ when no rain occurred.

Table 32-1. Data used in example evaluation of probability forecasts.

	Probability		Probability	
	RAIN		NO RAIN	
Occasion	Forecast (F_{ij})	Observed (E_{ij})	Forecast (F_{ij})	Observed (E_{ij})
1	0.7	0	0.3	1
2	0.9	1	0.1	0
3	0.8	1	0.2	0
4	0.4	1	0.6	0
5	0.2	0	0.8	1
6	0.0	0	1.0	1
7	0.0	0	1.0	1
8	0.0	0	1.0	1
9	0.0	0	1.0	1
10	0.1	0	0.9	1

The P-score is then given by:

$$\begin{aligned} P = (1/10)[&(0.7-0.0)^2+(0.9-1.0)^2+(0.8-1.0)^2+(0.4-1.0)^2 \\ &+(0.2-0.0)^2+(0.0-0.0)^2+(0.0-0.0)^2+(0.0-0.0)^2 \\ &+(0.0-0.0)^2+(0.1-0.0)^2+(0.3-1.0)^2+(0.1-0.0)^2 \\ &+(0.2-0.0)^2+(0.6-0.0)^2+(0.8-1.0)^2+(1.0-1.0)^2 \\ &+(1.0-1.0)^2+(1.0-1.0)^2+(1.0-1.0)^2+(0.9-1.0)^2]. \end{aligned} \quad (4)$$

The P-score in this example is 0.19, so the Brier score is 0.095 (half the P-score). This is a very low value, which indicates a high level of accuracy.

$$\begin{aligned} P = (0.1)(&0.49 + 0.01 + 0.04 + 0.36 + 0.04 \\ &+ 0.00 + 0.00 + 0.00 + 0.00 + 0.01 \\ &+ 0.49 + 0.01 + 0.04 + 0.36 + 0.04 \\ &+ 0.00 + 0.00 + 0.00 + 0.00 + 0.01). \end{aligned} \quad (5)$$

(2). The Brier score is not an ideal statistic for the evaluation of forecasts of rare events. In this case, a low value of the Brier score can be misleading. One can obtain a very good Brier score even when a rare event was never correctly forecast. To overcome this limitation, the Brier skill score can be used (Murphy and Winkler, 1982). The Brier skill score (BSS) evaluates the accuracy of a probability forecast in relation to a reference forecast, such as persistence or climatology. Mathematically, it is defined as

$$BSS = \left(1 - \frac{\text{Brier Score}}{\text{Brier Score of Reference Standard}}\right) \text{x } 100\%.$$

(3). If the reference forecast is climatology, then the Brier skill score measures how well a given set of forecasts improves over using climatology alone. Perfect forecasts earn a Brier skill score of 100 percent. Forecasts which are only as skillful as climatology receive a score of 0 percent, while forecasts that are inferior to climatology receive a negative skill score.

c. The Heidke Skill Score. The common skill score proposed by Heidke is defined as

$$S = \frac{R - E}{T - E} \quad (6)$$

where S is the Heidke skill score, R is the number of correct forecasts, T is the total number of forecasts, and E is the number of forecasts expected to be correct based on chance, climatology, or persistence. When calculating the Heidke skill score, a contingency table is normally set up as shown in Table 32-2.

Table 32-2. Contingency table showing the notation used to compute Heidke skill scores.

Observed	Forecast		
	OCCURRENCE	NON-OCCURRENCE	TOTAL
OCCURRENCE	a	b	x=a+b
NON-OCCURRENCE	c	d	y=c+d
TOTAL	m=a+c	n=b+d	T=a+b+c+d

Using the notation in the table, the Heidke skill score against chance (HSS) is given by:

$$HSS = \frac{(a + d) - \left(\frac{mx + ny}{T}\right)}{T - \left(\frac{mx + ny}{T}\right)}. \qquad (7)$$

(1). To illustrate how the HSS is calculated, consider the following example, which uses the data shown in Table 32-3.

Table 32-3. Contingency table with data used to compute Heidke skill score example.

Observed	Forecast		
	OCCURRENCE	NON-OCCURRENCE	TOTAL
OCCURRENCE	87	23	110
NON-OCCURRENCE	29	306	335
TOTAL	116	329	445

The following equation shows the Heidke skill score for this case:

$$HSS = \frac{(87 + 306) - \left[\frac{(116 \, x \, 110) + (329 \, x \, 335)}{445}\right]}{445 - \left[\frac{(116 \, x \, 110) + (329 \, x \, 335)}{445}\right]} = 0.69. \qquad (8)$$

(2). The Heidke skill score against chance has a value of one when all forecasts are correct, and a value of zero when the actual number of correct forecasts equals the number of correct forecasts that would occur by chance. It is possible to obtain a negative HSS when the forecasting method has no skill - it performs worse than random guessing. The Heidke skill score may be used to compare different forecast techniques. The technique having the largest HSS is determined to be the most useful.

(3). The Heidke skill score is not restricted to just two categories. For the case of R categories, the HSS is computed using the equation shown below:

$$HSS = \frac{\sum_{i=1}^{R} x_{ii} - \frac{1}{x_{TT}}\left(\sum_{i=1}^{R} x_{Ti} x_{iT}\right)}{x_{TT} - \frac{1}{x_{TT}}\left(\sum_{i=1}^{R} x_{Ti} x_{iT}\right)}. \qquad (9)$$

where x_{ij} is the number of observations with a forecast category i and observed category j. The subscript T refers to the totals column (x_{iT} is the total number of observations with forecast category i; x_{TT} is the total number of observations).

(4). Appleman (1960) felt the Heidke skill score was not measuring forecasts against a true standard. The standard used by the Heidke skill score, the expected number of correct forecasts based on pure chance, is given by the equation shown below (from equation 2).

$$E = \frac{(mx + ny)}{T} \qquad (10)$$

Appleman noted that this value is dependent upon the number of forecasts issued for each category. He maintains that the standard should be independent of the forecasts being evaluated. To overcome this problem, he proposed a new score. Appleman's score uses the same equation as the HSS (equation 1), but the value of E in the table is replaced with the number of observations in the category that is observed most frequently. That is

$$E = MAX(x, y). \qquad \textbf{(11)}$$

In the above example, E is be 335, resulting in an Appleman score of 0.52 (compared to 0.69 for the HSS).

(5). Murphy and Katz (1985) reviewed a number of skill scores that propose to measure the accuracy of categorical forecasts. They identified the following scores as being the best for measuring skill: Kuiper's performance index, Gringorten's skill score, Heidke skill score, Pierce's success index, and the Appleman score.

d. Secondary verification scores. The Heidke skill score is the most widely used of all the skill indices. Its greatest weakness is in the evaluation of forecasts dealing with rare events. Often, a large HSS can be obtained simply by never forecasting the rare event. The Appleman score also suffers from this deficiency. To supplement these skill scores, Goldsmith (1989) recommended the use of three secondary verification scores. The proposed secondary scores are the probability of detection (POD), the false alarm ratio (FAR), and the critical success index (CSI), which are described below.

(1). The probability of detection is simply the number of correct forecasts (x) of a given event divided by the total number of cases observed (T_o)

$$\text{POD} = \frac{x}{T_o}. \qquad \textbf{(12)}$$

(2). The false alarm ratio is the number of times the event was incorrectly forecasted to occur divided by the total number of the event forecasts (T_f)

$$\text{FAR} = \frac{(T_f - x)}{T_f}. \qquad \textbf{(13)}$$

(3). The critical success index is given by

$$\text{CSI} = \frac{x}{x + (T_f - x) + (T_o - x)}. \qquad \textbf{(14)}$$

32.3 Example. Suppose we want to measure the skill of a technique for forecasting precipitation type. Three categories of precipitation are to be forecast: freezing precipitation (Z), frozen precipitation (S), or liquid precipitation (R). The contingency table showing the results of this test is given below.

Table 32-4. Contingency table showing data for each precipitation category used in the examples.

Observed	Forecast			
	Z	S	R	TOTAL
Z	445	766	464	1675
S	445	18593	2420	21458
R	312	1673	30418	32403
TOTAL	1202	21032	33302	55536

The Heidke skill score against chance for this example is 0.782. This suggests a technique with considerable skill. To better understand the strengths and weaknesses of this forecasting method, the secondary verification scores are computed for each forecast category.

Table 32-5. Forecast verification measures.

	Z	S	R
POD	0.27	0.87	0.94
FAR	0.63	0.12	0.09
CSI	0.18	0.78	0.86

These scores point out that the large HSS is misleading. The calculation is dominated by the large number of the liquid (R) and frozen (S) forecasts. Freezing precipitation has low probability of detection and a high false alarm rate. This results in a poor critical success index. Thus, despite the model's very high HSS, it displays poor performance in forecasting freezing precipitation (Z).

32.4 Conclusion. Despite the large number of skill scores which have been proposed, no single score is clearly superior. It is perhaps unrealistic to expect one number to describe the many facets of forecast verification. The Heidke skill score against chance is the most widely used measure for categorical forecasts. However, it can provide misleading results, especially when evaluating forecasts of relatively rare events. Following Goldsmith (1989), we recommend supplementing the HSS with the probability of detection, false alarm ratio, and the critical success index. These secondary scores can identify situations when the HSS is subject to bias.

Chapter 33

MODEL EVALUATION

33.1 Introduction. Model evaluation involves the comparison of model predictions with observations. The root mean square error (RMSE) is often used at AFCCC as an accuracy measure. This chapter discusses the RMSE and related statistics.

33.2 Discussion.

a. The equation shown below is used to calculate the RMSE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2} \qquad \textbf{(1)}$$

where: o_i and p_i are individual elements of the observed and predicted distributions, and N is the total number of data pairs. The closer the RMSE is to zero, the better the fit between observed and predicted values.

b. The mean absolute error (MAE) is calculated by the equation shown below.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|p_i - o_i|. \qquad \textbf{(2)}$$

c. According to Willmot (1982), RMSE and MAE are among the best overall measures of model performance, but both have their limitations. Graphical data displays such as scatterplots, box plots, histograms, or displays of cumulative frequency distributions should always accompany the RMSE or MAE. These displays help identify patterns within the errors and extreme cases. The correlation coefficient is not recommended as a measure of accuracy, as it is often misleading (Panofsky, 1958).

d. Segal and Pielke (1981) identified two conditions that must be satisfied for a model to demonstrate skill:

(1) The standard deviation of the observed values must be similar to the standard deviation of the predicted values.

(2) The RMSE divided by the standard deviation of the observed values must be less than one.

e. Fox (1981) recommended using the following test statistic to evaluate model performance:

$$\bar{d} \pm t_{.05}\sqrt{\frac{s_o^2}{n_o} + \frac{s_p^2}{n_p}} \qquad \textbf{(3)}$$

$$s^2 = \frac{(n_o - 1)s_o^2 + (n_p - 1)s_p^2}{n_o + n_p - 2}$$

$$\bar{d} \pm t_{.05}\sqrt{\frac{2s^2}{n}}$$

where: $\bar{d}$ is the difference between the observed mean and the predicted mean; s_o^2 is the variance of the observed distribution; s_p^2 is the variance of the predicted distribution; n is sample size and $t_{.05}$ is the value of the t-statistic for a 95-percent confidence level with (2n - 2) degrees of freedom. Test statistic assumes equal sample sizes and the population variances are the same. Equation 3 tests the null hypothesis that there is no statistical difference between the mean observed and the mean expected values. If zero lies within the confidence interval specified by equation 3, than one accepts the null hypothesis. If zero lies outside this interval, the null hypothesis is rejected.

f. Bias or mean error (ME) is simply the difference between the average forecast and average observation, and therefore expresses the bias of the forecasts.

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i). \qquad (4)$$

Forecasts that are on the average too high will exhibit ME>0, and forecasts that are on the average too low will exhibit ME<0. It is important to note that the bias gives no information about the typical magnitude of individual forecasts, and is therefore not an accuracy measure.

33.3 Example. The following example will help to illustrate the points raised above.

a. Suppose one fits a set of temperature observations to a model, and wishes to test the closeness of fit. Table 33-1 shows the observed and predicted values. Table 33-2 shows the means and standard deviations.

b. The root mean square error for this example is

$$RMSE = \sqrt{\frac{36.11}{8}} = 2.12. \qquad \textbf{(5)}$$

The standard deviation of observed (22.1) and modeled (21.9) distributions are similar. The RMSE divided by the standard deviation of observed values is 0.096. Since this is less than one, the model has demonstrated skill according to the criteria of Segal and Pielke (1981).

c. Next, the analyst will use the statistic recommended by Fox (1981) as noted in equation 3:

$$0.60 \pm 2.145\left(\sqrt{\frac{(2)(484.01)}{8}}\right) \qquad \textbf{(6)}$$

$$\text{where } s^2\,(484.01) = \frac{(7)(488.41)+(7)(479.61)}{8+8-2}$$

$$0.60 \pm 23.60.$$

The upper limit of this interval is 24.20 and the lower limit is -23.00. Since zero lies within this interval, the analyst concludes the modeled distribution is a good fit to the observed data.

Table 33-1. Observed and predicted temperature values.

Temperature Observations	Predicted Temperatures	Differences
12.4	11.9	0.5
24.3	21.8	2.5
35.2	37.9	-2.7
41.1	40.6	0.5
50.6	50.5	0.1
64.2	60.3	3.9
66.2	68.1	-1.9
76.1	74.3	1.8

Table 33-2. Mean and standard deviation of observed and predicted temperature values.

	OBSERVED	PREDICTED
MEAN	46.3	45.7
STD DEV	22.1	21.9

BIBLIOGRAPHY

Air Weather Service, *Guide for Applied Climatology*, AWS-TR-77-267, 1977.

Air Weather Service, *Some Techniques for Deriving Objective Forecasting Aids and Methods*, AWS Technical Report 235, 1955.

Air Weather Service Technical Report 105-25, *Study of Length of Record Needed to Obtain Satisfactory Climatic Summaries for Various Meteorological Elements*, AWS, 1943.

Afifi, A.A., and V. Clark, *Computer-Aided Multivariate Analysis*, Van Nostrand, Reinhold Co., 1984.

Appleman, H.S., "A Fallacy in the Use of Skill Scores," *Bull. Amer. Met. Soc.*, 41, 64-67, 1960.

Bevington, P. R., and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, Second Edition, McGraw-Hill, 1992

Blom, G., "Extrapolation of Linear Estimates to Larger Sample Sizes," *Journal of American Statistical Association*, Vol. 75, pp.912-917, 1980.

Boehm, A., *Transnormalized Regression Probability*. AWS/TR—75-259, 1976.

Boehm, A., J. Willand, and G. Gray, *"Reconciling Satellite and Surface Cloud Observations," Eighth Conference on Applied Climatology*, 1993.

Bradley, J.V., "The Central Limit Effect for a Variety of Populations and the Influence of Population Moments," *Journal of Quality Technology*, Vol 5, pp. 171-177, 1973.

Brier, G.W, "Verification of Forecasts Expressed in Terms of Probability," *Mon Wea Rev*, 78, pp. 1-3, 1950.

Brooks, C.E.P. and N. Carruthers, *Handbook of Statistical Methods in Meteorology*, Her Majesty's Stationary Office, London, 1953.

Burroughs, L.D., "Forecasting Open Ocean Fog and Visibility," *11th Conf. on Probability and Statistics*, 1989.

Chu, P., and Y. He, "Prediction of Hawaiian Winter Rainfall Using Canonical Correlation," *12th Conference on Probability and Statistics in the Atmospheric Sciences, Toronto, Ontario, Canada,* 1992.

Coffin, C.R., and A.J. Warren, *Zaragoza AB Fog Study*, USAFETAC/PR-91/015, 1991.

Coffin, C.R., *Background Paper on Discriminant Analysis.* USAFETAC, Scott AFB, Ill., September, 1991.

Cornell, Daniel, *Thunderstorm Forecast Study for Eglin AFB, Fla.* USAFETAC PR-93/001, Scott AFB, Ill., March, 1993.

Court, A., *Climatic Normals as Predictors, (Report Numbers 67-82-1 and 68-82-5)*, Prepared for Air Force Cambridge Research Laboratories, Office of Aerospace Research, United States Air Force, Bedford, Mass., 1967-1968.

D'Agostino, R.B., A. Belanger, and R.B. D'Agostino, Jr., "A Suggestion for Using Powerful and Informative Tests of Normality, *The American Statistician*, Vol. 44, pp 316-321, 1990.

David, H.A., *Order Statistics*, Second Edition, New York: Wiley, 1981.

Draper, N.R., and H. Smith, *Applied Regression Analysis*, Second Edition, John Wiley & Sons, 1981.

Fox, D.G., "Judging Air Quality Model Performance," *Bull. Amer. Meteor. Soc.*, Vol 62, pp. 559-609, 1981.

Goldsmith, B.S., "A Comprehensive Analysis of Verification Results for Forecasts of Precipitation Type and Snow Amount." *Paper presented at the 11th Conf. on Probability and Statistics*, 1989.

Hair, J. F., R. E. Anderson, and R. L. Tatham. *Multivariate Data Analysis*, Second Edition, MacMillan Publishing Company, 1987.

Hays, W.L., and R.L. Winkler. *Statistics: Probability, Inference, and Decision*, New York: Holt, Rinehart and Winston, Inc., 1971.

Hosking, J. R. M., FORTRAN Routines for Using the Method of L-moments, Version 2, *IBM Research Report*, RC17097, IBM Research Division, T. J. Watson Center, Yorktown Heights, New York, 1991.

Hosking, J. R. M., "L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics," *Journal of Royal Statistical Society*, Vol. 52, No. 1, pp. 105-124, 1990.

Hosking, J. R. M. and Wallis, J. R., "Some Statistics Useful in Regional Frequency Analysis," *Journal of Water Resources Research*, Vol. 29, No. 2, pp. 271-281, 1993.

Jagannathan, P., Arlery, R., TenKate, H., and Zavarina, M.V., *WMO Technical Note 84*. WMO - No. 208.TP.108, 1967.

Kenney, J.F. and E.S. Keeping, *Mathematics of Statistics*, D. Van Nostrand Company, 1957.

Klecka, W.R., *Discriminant Analysis*, Sage Publications, 1980.

Kruizinga, S., K. Kok, and L. Wilson, "The WMO Training Workshop on the Interpretation of NWP Products in Terms of Local Weather Phenomena and Their Verification," *12th Conference on Probability and Statistics in the Atmospheric Sciences, Preprint.* Toronto, Ontario, Canada, Jun 1992.

Landsberg, H.E. and W.C. Jacobs, *Applied Climatology, Compendium of Meteorology*, American Meteorological Society, 1951.

Larson, D.A. (1994), "Adding Guidelines for Influence Diagnostics to Output from the Reg Procedure," *Observations: The Technical Journal for SAS Software Users*, Vol 3, No 2, pp 54-60.

Law, A.M. and W.D. Kelton, *Simulation Modeling and Analysis, Second Edition*, McGraw-Hill, 1991.

Littell, R.C., R.J. Freund, and P.C. Spector, *SAS Systems for Linear Models*, Third Edition, SAS Institute, Cary N.C., 1991.

McCutchan, M.H., and M.J. Schroeder, "Classification of Meteorological Patterns in Southern California by Discriminant Analysis," *J. Appl. Meteor.,* Vol. 12, pp. 571-577, 1973.

Miller, R.G., "Statistical Prediction by Discriminant Analysis," *Meteorological Monographs,* Vol 4, No 25, 1962.

Murphy, A.H. and Katz, R.W., *Probability, Statistics, and Decision Making in Atmospheric Sciences*, Westview Press, 1985.

Murphy, A.H., and R.L. Winkler, "Subjective Probabilistic Tornado Forecasts: Some Experimental Results," *Mon Wea Rev*, 110, pp. 128-1297, 1982

Myers, R.H., *Classical and Modern Regression With Applications*, PWS-Kent, Boston, Mass., 1990.

Neter, J., W. Wasserman, and M. Kutner, *Applied Linear Statistical Models*, Irwin, Calif., 1990.

Netter, J., Wasserman, W., and Kutner, M.H., *Applied Linear Statistical Models*, Irwin, Calif., 1985.

Panofsky, H.A., and G.W. Brier, *Some Applications of Statistics to Meteorology*, Pennsylvania State University Press, University Park, Pa., 1958.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes*, Cambridge University Press, 1986.

Sachs, L., *Applied Statistics - A Handbook of Techniques*, Springer - Verlag, 1984.

SAS, *SAS Procedures Guide Version 6, 3rd Edition.* SAS Institute Inc., Cary, N.C., 1990.

SAS. *SAS Users Guide: Statistics*, Version 5. SAS Institute Inc., Cary, N.C., 1985.

SAS, *Categorical Data Analysis Course Notes*, SAS Institute, Inc., Cary, N.C., 1992.

SAS, *SAS Technical Report*, Release 6.07. SAS Institute Inc., Cary, N.C., 1992.

SAS, *SAS System for Linear Models*, 3rd Edition, SAS Institute, Cary, N.C., 1991.

SAS, *SAS/ETS Users Guide*, Version 5. SAS Institute Inc., Cary, N.C., 1984.

SAS, *SAS/ETS Users Guide*, Version 6, SAS Institute, Inc., Cary, N.C., 1988.

SAS, *SAS/STAT Users Guide* - Version 6, 4th Edition, Vol. 1. SAS Institute, Cary, N.C., 1990.
Schlotzhauer, S.D., and R.C. Littel, *SAS® System for Elementary Statistical Analysis*, SAS Institute Inc., 1987.

Schulman, R. S., *Statistics in Plain English, With Computer Applications.* Van Nostrand Rheinhold, 1992.

Searle, S.R., *Linear Models*, New York: John Wiley and Sons, 1971.

Segal, M. and R.A. Pielke, "Numerical Model Simulation of Human Biometeorological Heat Load Conditions: Summer Day Case," *J. Appl. Meteor.*, Vol 20, pp. 735-749, 1981.

Snedecor, G.W., and W.G. Cochran, *Statistical Methods*, Seventh Edition, Iowa State University Press, 1980.

Steel, R., and J. Torrie, *Principles and Procedures of Statistics*, McGraw-Hill, 1980.

Thompson, M.L. and W. Zucchini, "Assessing the Value of Probability Forecasts," *Mon. Wea. Rev.*, Vol 118, pp. 2696-2706, 1990.

Vardeman, S., "What About Other Intervals?", *The American Statistician*, August 1992.

Weisberg, S., *Applied Linear Regression*, Second Edition. John Wiley and Sons, 1985.

Whiton, R.C., E.M. Bertecek. *Basic Techniques in Environmental Simulation,* USAFETAC/TN-82/004, United States Air Force Environmental Technical Applications Center, 1982.

Willand, J. H., *Database Blending for the Climatology of Cloud Statistics Program.* PL-TR-92-2344, 1992.

Willmot, C.J., "Some Comments on Model Performance," *Bull. Amer. Meteor. Soc.,* vol 63, pp. 1309-1313, 1982.

World Meteorological Organization, WMO-NO.208.TP.108: *A Note on Climatological Normals*, WMO, Geneva, 1967.